DEVELOPMENT AND EVALUATION OF PROTEIN DESIGN METHODS FOR FUNCTIONAL TARGETS

Thesis by

Christina Luisa Vizcarra

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2008

(Defended May 23, 2008)

Acknowledgements

It has been a privilege to be at Caltech for the last six years. I thank my advisor, Steve Mayo, for providing a great work environment and for allowing me to learn so many new techniques. Even if it was not in the interest of getting research done quickly, the opportunity to participate in both computational and experimental projects certainly furthered my education. I am grateful to the members of my committee, Doug Rees, Frances Arnold, and Jim Heath for advice throughout this process. I am also grateful to the funding sources that provided support for my graduate work: the Rosen fellowship from Caltech and the National Science Foundation's graduate research fellowship.

The Mayo lab has been a unique work environment that I feel fortunate to have been a part of. I am thankful to Ben Allen for providing advice and more computer help than should be asked of one graduate student. Ben and Possu Huang were supportive and helpful baymates. Roberto Chica joined the group near the end of my time here but has been a wonderful presence in the lab. It was both enjoyable and educational to work with my collaborators in the lab Corey Wilson, Tom Treynor, Daniel Nedelcu, and Shannon Marshall. I owe much to the other members of the Mayo Lab, in order of appearance: Premal Shah, Geoffry Hom, Rhonda DiGiusto, Scott Ross, Marie Ary, Cynthia Carlson, Peter Oeschlager, J.J. Plecs, Julia Shifman, Jonathan Kyle Lassila, Jessica Mao, Eun Jung Choi, Eric Zollars, Oscar Alvizo, Jennifer Keeffe, Heidi Privett, Karin Crowhurst, Barry Olafson, Cathy Miles, Alex Nisthal, Erin Drez, Matthew Moore, Swathi Adindla, and Kurt Mou. Many other good friends at Caltech made my time here great: Justin Bois, John Keith, Jesse Bloom, Bonnie Sheriff, Amie Boal, and Heather Wiencko.

I have had the opportunity to work with a great group of collaborators. Ned Wingreen, Chen Zeng, and Naigong Zhang were thoughtful and always responsive. My time working with Mike Chen and Chris Snow in the Arnold lab was a lot of fun. I received a great deal of help from Emil Alexov and Barry Honig at the start of my projects.

I am particularly grateful to past educators who have helped me get to this point: Dave Benson, Estela Gavosto, and Ward Thompson at the University of Kansas, and Cole Ogdon at Shawnee Mission East High School.

My family has provided me with constant support and love. I thank my parents Mary and Jorge Vizcarra for always encouraging me to make education a central focus of my life. I thank my many siblings, nieces, nephews, cousins, and in-laws for being supportive and for visiting us out here on the west coast. Finally, I owe the most thanks to my husband Matthias, who moved to L.A. with me and makes life beautiful.

Abstract

Computational protein design seeks to identify amino acid sequences that will fold into a specified three-dimensional structure. Extending this technique from identification of sequences that retain a native structure to the design of sequences that will carry out a function has been a significant challenge. Modeling the energetics of catalysis and binding requires considerations that may not be necessary for the design of folded, stable proteins. I have investigated models for protein electrostatics with the goal of improving current methods for the design of functional molecules. The work in this thesis is focused on the Poisson-Boltzmann model, a dielectric continuum model that describes the effect of solvent polarization on the electrostatic potential in a protein. I found that this model is amenable to design calculations, as judged by its ability to be decomposed into terms that are used in sequence selection.

Aside from energy estimation, there are a number of assumptions that are made in protein design in order to make the problem computationally tractable. Because of these assumptions, and also because of incomplete models of protein function, it is expected that many proteins sequences will need to be experimentally characterized to find one that meets a difficult design goal. To this end, I examined methods for using computational tools to produce libraries of protein sequences. These studies showed that (1) structure-based, computational library design methods can be used to generate libraries with a high number of folded proteins and (2) computational design is a promising tool for generating highly mutated proteins with a diverse range of functions.

TABLE OF CONTENTS

FIGURES AND TABLES

ABBREVIATIONS

AA	amino acid
AUC	analytical ultracentrifugation
BM-3	cytochrome p450 from *Bacillus megaterium* (accession # P14779)
C	consensus
CD	circular dichroism
CFA	Coulomb field approximation
CPK	color scheme based on Corey, Pauling, and Kultun models
DBIS	diversity benefit applied to interacting sets
DDD	distance-dependent dielectric
DEE	dead-end elimination
DME	dimethyl ether
DNA	deoxyribonucleic acid
DTT	dithiothreitol
E. coli	*Escherichia coli*
ENH	engrailed homeodomain
epPCR	error-prone PCR
FASTER	fast and accurate sidechain topology and energy refinement
FDPB	finite-difference Poisson-Boltzmann
G0	no generic sidechains
G3	three-sphere generic sidechain
GB1	B1 domain of Protein G
GB	generalized Born
GBSA	generalized Born/surface area
GB-PDA	generalized Born with pairwise descreening approximation
GFP	green fluorescent protein
GMEC	global minimum energy conformation
HPLC	high performance liquid chromatography
IPTG	isopropyl β-D-1-thiogalactopyranoside
LK	Lazaridis-Karplus (solvent exclusion model)
MC	Monte Carlo
MSA	multiple sequence alignment
NSC	negatively supercharged ENH variant
Ni-NTA	nickel-nitrilotriacetic
ORBIT	optimization of rotamers by iterative techniques
PB	Poisson-Boltzmann
PDB	protein data bank
PCR	polymerase chain reaction
PNK	polynucleotide kinase
RMSD	root mean squared difference
SE	site entropy
SCMF	self-consistent mean field
TFA	trifluoroacetic acid
VDW	van der Waals
WT	wild type

Chapter 1

Introduction

The central goal of computational protein design is to identify amino acid sequences that will fold into a given three-dimensional structure.[1,2] This is accomplished using the scheme shown in Figure 1-1. The design process starts with the selection of a three-dimensional protein structure. Residues to be designed are selected: this might include all of the residues in the protein or just a subset as shown in Figure 1-1. The conformational flexibility of the candidate amino acid sidechains at each position are modeled using discrete conformations, referred to as rotamers.[3] The energy of each rotamer is calculated using an energy function that is primarily based on molecular mechanics force fields.[4,5] Energies are stored for each rotamer's interaction with the rest of the protein and for the interaction between all pair of rotamers. This energy table and specialized search algorithms are used to search the multidimensional sequence/energy landscape to find the optimal rotameric sequence.[6] A number of variations have been made on the general computational protein design scheme in Figure 1-1.

Recently, the goal of protein design has expanded from retention of a target fold to include the design of novel function.[7–10] This is accomplished by modeling the structure in a functionally relevant state and designing an amino acid sequence that will

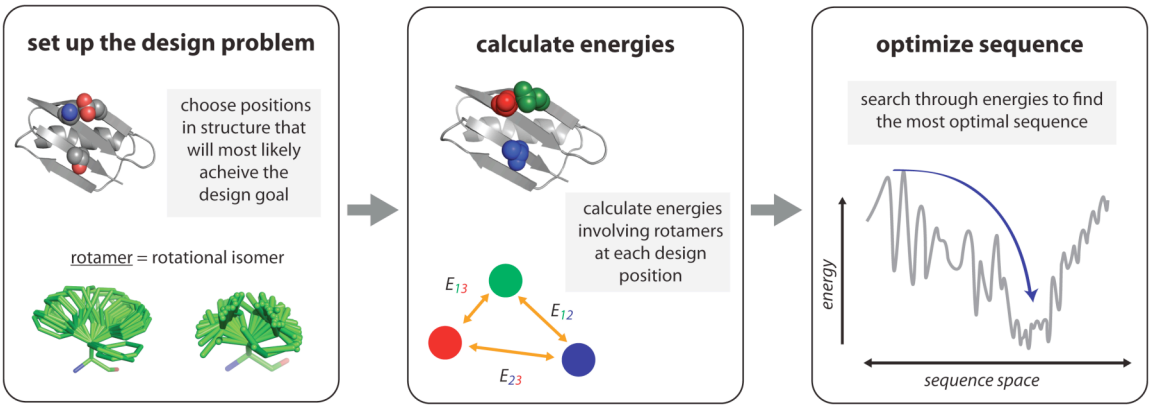


Figure 1-1. Computational protein design. An example is shown in which three positions on the surface of protein G are designed.

favor that conformational or chemical state. The design of functional molecules has presented a formidable challenge for the field of computational protein design. Specifically, binding and catalysis have been refractory to *de novo* design. We do not have clear answers for why these are such difficult targets, but one can look at the many assumptions in the process outlined in Figure 1-1 to compile a list of what factors might hinder the design of function:

1. Crude approximation of energies
2. Fixed protein backbone and other limitations on conformational sampling
3. Lack of consideration of multiple states
4. Poor models for the relevant chemical states.

The central goal of my graduate work was to address the first issue. To this end we investigated models for polar interactions, an energetic contribution expected to be crucial for catalysis and binding. The residues involved in catalysis are overwhelmingly polar.[11,12] Similarly, the residues found in protein-protein interfaces more closely resemble the composition of protein surfaces rather than protein cores.[13] Therefore, both of these processes will require accurate modeling of the balance between favorable electrostatic interactions and the energetic cost of desolvating polar groups. I pursued my thesis work with the idea that improved modeling of electrostatic interactions will accelerate the design of functional molecules.

Chapter 2 of this thesis outlines in detail the many methods that have been used by protein designers to account for electrostatic interactions. All of the models reported thus far are constrained by the enormous computational demands of the protein design problem. For example, in a modestly sized design calculation described later in this

thesis, more than 10^{61} possible rotameric sequences are possible. A one second calculation to model the energies of electrostatic interactions in each sequence conformation would lead to a total calculation time of 15762426552055500000000000000000000000000000000000000000 years. Because of this challenge, all energies are calculated in a "residue pairwise" scheme in which the energy only reflects the interaction energy between two rotamers ("two-body") or between a rotamer and the rest of the protein that is not being designed ("one-body"). Search algorithms can then use this table of one-body (E_i) and two-body (E_{ij}) energies to calculate sequence energies as needed

$$E = \sum_{i=1}^{n} E_i + \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} E_{ij} .$$

The computational benefit of using a residue pairwise calculation comes at the cost of accurate modeling of the energies since our most complete theories of protein energetics have "many-body" terms.

Chapters 3 and 4 describe work on formulating a Poisson-Boltzmann (PB) model that can be implemented in current protein design protocols. The PB model is a continuum solvation model in which the dielectric environment and the charge distribution of the protein determine the electrostatic potential at each atom in the protein. Since the potential is dependent on the dielectric environment, which is itself dependent on the position of all atoms in the protein, the PB is a many-body energy model. We address the issue of whether the one- and two-body energy terms used to guide sequence selection would be meaningful if derived from the PB model. It was found that the one- and two-body energy terms provide PB energies that are similar to the energy of the

standard many-body PB model, indicating that the PB model is potentially useful for protein design methods. In Chapter 3, the initial pairwise formulation is introduced, and in Chapter 4, improvements to this formulation plus comparison with additional models are presented. I also investigated the Generalized Born (GB) model as an alternative to the computationally expensive PB model. In Appendix B, I show that the accuracy of the GB model, judged as the ability to reproduce PB energies, is similar to models that are currently used in ORBIT. I identify features of the GB model that might lead to its insensitivity to the microenvironments that are sampled in a protein design calculation.

The promising computational results from Chapters 3 and 4 lead me to implement the residue pairwise PB model into ORBIT and assess the validity of this model in experimental tests. In Chapter 5, data is presented for these efforts and for computational characterization of the PB model's treatment of hydrogen-bonded sidechains. I used the design of the surface residues in *Drosphilia melanogaster* engrailed homeodomain (ENH) as an experimental test case. Using the PB model from Chapter 4 did not lead to the design of a stabilized variant of ENH. In attempting to make a comparison between the PB-designed sequences and those designed with other energy functions, I found unexpected behavior in this test case. This behavior highlights the fact that design calculations are highly sensitive to factors that might be unrelated to the energy function. In this chapter, I also investigate the problem of reconciling the incomplete description of hydrogen-bonding inherent in continuum solvation, an issue that extends beyond protein design calculations and must be addressed for enzyme design. In a computational experiment on a set of crystallographic, hydrogen-bonded sidechain pairs, I showed that

the rotamers chosen by the PB model do not necessarily conform to the geometric description of hydrogen bonds used currently in the ORBIT force field.

During the experimental characterization discussed in Chapter 5, I made a series of ENH variants to dissect the relationship between rotamer library and energy function. One of these variants, designed using the standard ORBIT force field and sequence biasing, had a melting temperature around 95°C. Chapter 6 discusses characterization of this variant and also a "supercharged" variant. These two molecules represent extremes of surface plasticity for the ENH fold: one has highly optimal surface electrostatic properties while the other has a high degree of repulsion between its surface residues. I also investigated the role of surface electrostatics in WT protein G. In Appendix A, data is presented for an ion pair on the beta-sheet surface protein G. This ion pair was found to have a favorable free energy of interaction, but removing it caused a negligible change in the protein's thermodynamic stability.

An alternative or complementary approach to addressing the shortcomings listed at the beginning of this chapter is to take a higher throughput strategy in tackling difficult design targets. Where possible, medium to high throughput screening could be used to characterize many computationally designed molecules. The challenge is turning the information from the design calculation into combinatorial libraries of sequences that can be synthesized in the laboratory. Chapter 7 discusses the evaluation of a number of different library design strategies by their ability to create libraries that (1) retain function in the largest number of library members and (2) perturb the function of the WT parent. Using green fluorescent protein as a test case, we defined the retention of function as some measurable threshold of fluorescence, which itself could be considered a lower

bound on the number of folded sequences in the library. Diversity of function was defined by shifts in the emission peak position. Our experiments showed that structure-based design methods perform well by both metrics: retention and diversity of function. In Appendix C, data is presented for the design of combinatorial libraries focused on the substrate-binding pocket of cytochrome p450 BM3 from *Bacillus megaterium*. The p450 libraries were screened in the laboratory and shown to have a high number of folded variants.

The sum of the work in this thesis is the development and evaluation of computational tools that can be used in the design of functional molecules. The current outlook for designing enzymes and binding proteins is positive. Recently, much progress has been made in the field of ligand placement in active sites[14,15] and impressive strides have been made in designing enzymes *de novo*.[9,10] There is also progress in the field protein-protein interaction design.[16,17] With improved modeling strategies and hybrid engineering methods, the technological benefits of protein-based devices and catalysts will be realized.

References

1. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.

2. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87.

3. Dunbrack, R. L., Jr. (2002). Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440.

4. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513.

5. Mendes, J., Guerois, R. & Serrano, L. (2002). Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **12**, 441–446.

6. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299**, 789–803.

7. Bolon, D. N. & Mayo, S. L. (2001). Enzyme–like proteins by computational design. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14274–14279.

8. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185–190.

9. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro–aldol enzymes. *Science* **319**, 1387–1391.

10. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., Dechancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195.

11. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121.

12. Gutteridge, A. & Thornton, J. M. (2005). Understanding nature's catalytic toolkit. *Trends Biochem. Sci.* **30**, 622–629.

13. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.

14. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. (2006). Combinatorial methods for small–molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16710–16715.

15. Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Rothlisberger, D. & Baker, D. (2006). New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **15**, 2785–2794.

16. Huang, P. S., Love, J. J. & Mayo, S. L. (2005). Adaptation of a fast Fourier transform–based docking algorithm for protein design. *J. Comput. Chem.* **26**, 1222–1232.

17. Huang, P. S., Love, J. J. & Mayo, S. L. (2007). A de novo designed protein protein interface. *Protein Sci.* **16**, 2770–2774.

Chapter 2

Electrostatics in computational protein design

The text of this chapter is adapted from a published review article that was co-authored with Professor Stephen L. Mayo

C.L. Vizcarra and S.L. Mayo, *Current Opinion in Chemical Biology* **9**, 622–626 (2005).

Abstract

Catalytic activity and protein-protein recognition have proven to be significant challenges for computational protein design. Electrostatic interactions are crucial for these and other protein functions, and therefore accurate modeling of electrostatics is necessary for successfully advancing protein design into the realm of protein function. This review focuses on recent progress in modeling electrostatic interactions in computational protein design, with particular emphasis on continuum models.

Introduction: the electrostatics challenge

Computational protein design seeks to design the amino acid sequence of a protein in a manner that preserves the target three-dimensional fold.[1,2] The compatibility of an amino acid sequence with the target fold is determined by an energy function. Standard components of protein design energy functions are van der Waals, solvation, electrostatics, hydrogen bonding, and various statistical terms that approximate entropy and other forces that are not modeled explicitly.[3,4] The balance between these energetic terms has generally been trained on experimental stability data[1,5] or on the ability to recover wild-type amino acid composition.[6,7]

Recently, the goals of many computational protein design projects have shifted from preserving the folded structure to designing function. Electrostatic interactions play important functional roles in many biomolecular systems. In enzymes, surface electrostatic potential can channel substrates to the active site,[8] where the electrostatic environment plays a key role in stabilizing the transition state.[9] Because enzymes are such efficient catalysts, the *de novo* design of enzymatic activity has many technological applications.[10] Protein-protein interfaces contain a proportion of polar and charged residues similar to that on the protein surface,[11] and therefore their design requires a careful balancing of polar desolvation energy and electrostatic interactions.[12–14] The control of protein recognition is an important goal for protein designers, as this will allow for the manipulation of biochemical networks[15] in ways that may shed light on signal transduction mechanisms and potentially lead to the design of novel biological circuits.

Residues that impart function may compromise stability.[16] The degree to which electrostatic interactions stabilize the folded state of a protein has been the subject of

much debate.[17] It has been suggested that electrostatic interactions may play a role in fold specificity instead of stability.[18] Since electrostatics may have a relatively small net contribution to the free energy of folding in most mesophilic proteins, one can design well-folded, stable proteins by focusing on producing well-packed, hydrophobic cores and using only a very crude or damped model for electrostatic interactions.[19] Indeed force fields with an orientation-dependent hydrogen bonding potential and a small or non-existent Coulombic term have yielded stable, well-folded proteins.[20–22] However, it stands to reason that a physical model that accurately captures the electrostatic forces that allow a protein to fold should be adaptable to the challenges imposed by the desire to design protein function.

Marshall et al.[23] showed that for the surface of an all alpha-helical protein, current electrostatic models used in computational protein design did not accurately capture the electrostatic effects of helix dipole and N-capping interactions. Restricting the amino acid identities at N-cap positions to those that have high N-capping propensities and restricting the charge of amino acids at the N-terminal and C-terminal regions of the helix allowed for the design of a sequence that was stabilized by 3 kcal mol^{-1} over an unbiased design. Similarly, it has been shown that polar amino acids are found in the cores of natural proteins.[24,25] Bolon et al.[26] designed a stabilized variant of thioredoxin by imposing empirical hydrogen-bonding rules that would compensate for the cost of polar desolvation. In order to make protein design force fields more general, it is desirable to capture the balance between desolvation and electrostatic interaction energy through physical modeling as opposed to the heuristics used in the approaches described above.

In the simplest estimate, the number of sequences considered for even a small, 50-amino-acid protein is astronomically large (~ 10^{65} sequences). Most successful computational protein design algorithms approach this combinatorial problem by using computationally tractable pairwise decomposable energy functions that score the arrangement of at most two sidechain conformations at a time. The limitation to pairwise decomposable energy functions has led to the development of efficient sequence optimization algorithms[27] but has precluded certain energy models that do not lend themselves to pairwise expressions. Because proteins are surrounded by water, which is highly polarizable, any accurate description of protein electrostatics is a function of the solvent environment, making the electrostatic energy a many-body term. It is therefore necessary to reconcile the limitations of the pairwise approximation with the need for an accurate description of electrostatics. Furthermore, modeling of water explicitly is currently intractable for the number of conformational energies that must be calculated for protein design. Therefore continuum or empirical models have been used in most protein design force fields to address electrostatic interactions as well as polar desolvation. It has been pointed out by Jaramillo and Wodak[28] that protein design may be a stringent test of continuum models because design requires that an energy function distinguish between many micro-environments inside the protein.

In the past decade, great effort has gone into updating the electrostatics and polar solvation portions of molecular mechanics force fields. In this review, we focus on the advances in continuum electrostatics for computational protein design. It should be noted that protein design energy functions have often treated the desolvation of polar and charged sidechains as a separate term from the electrostatic interaction energy. Since

both of these terms are functions of the dielectric environment, continuum models generally propose one consistent treatment for solvation and electrostatics. We therefore consider the modeling of polar and charged residue desolvation as part of the electrostatics challenge. Previous reviews of computational protein design have covered general methodology,[29–31] energy functions,[3,4] protein-protein interactions,[32] metal centers,[33–35] and catalysis.[10] In addition, continuum models for electrostatics and solvation have been reviewed extensively.[36–38]

Working models

Poisson-Boltzmann The Poisson-Boltzmann (PB) equation is considered the standard for accuracy within the limitations of the continuum description. In PB calculations the solute is described as a low dielectric cavity embedded in a high dielectric solvent, and the induced polarization in the solvent is used to calculate the electrostatic potential at all points in the protein. Analytical solutions to the PB equation exist only for simple solute geometries such as spheres or cylinders. Numerical methods must be employed for complex shapes like that defined by a protein molecular surface.[8] Although the PB model is not readily pairwise decomposable by side chain, Marshall et al.[39] recently proposed a two-body formulation using the finite difference PB solver DelPhi. In the two-body approach, a reduced representation of the protein is used and perturbations to the dielectric boundary are considered explicitly for each sidechain conformation. Surprisingly, the energies produced by summing two-body terms are quite close to those obtained by calculation with the entire surface represented. A pairwise model in which all possible sidechain conformations are used to define the dielectric boundary has been

used by Georgescu et al.[40] for pKa calculations and may be useful in reducing the complexity of protein design calculations.

Modified Tanford-Kirkwood The original Tanford-Kirkwood model [41] treated proteins as spheres, allowing for an analytical solution to the PB equation.[42] Because advances in structural biology have shown the spherical representation to be a dubious approximation for many proteins, Havranek and Harbury[43] developed the modified Tanford Kirkwood (MTK) method in which the charge distribution of the protein is mapped from the exact protein geometry onto a sphere. They also use a shell charge representation and an image charge solution to calculate the electrostatic free energy associated with a protein conformation. This model, along with a negative design scheme, was used to create a series of coiled coil systems that specifically formed homo- or hetero-dimers.[44]

Generalized Born The Generalized Born (GB) model maps each charge in the protein to the center of a sphere with a radius that reflects the burial of the charge in the protein. From this Born radius, the electrostatic potential can be calculated at each charge in the protein.[45] GB implementations are distinguished by the method in which Born radii are computed. The GB model has enjoyed wide use in MD simulations, but it has been shown that the ability of the GB model to reproduce more accurate PB calculations varies widely between implementations.[46] It has also been shown that many commonly used GB methods drastically underestimate the burial of atoms in the protein core,[47] making GB a particularly insensitive model for scoring desolvation energies of individual polar and charged sidechains [Vizcarra and Mayo, unpublished results]. Pokala and Handel[48]

have adapted the GBSA method[49] for use in protein design. They overcome the pairwise decomposability problem by calculating Born radii using spheres to approximate sidechains of unknown identity.

Empirical Models Instead of attempting to calculate the screening of electrostatic interactions using Born radii or solvent polarization charges, a distance dependent dielectric model (DDD) has been used to damp all Coulombic interactions regardless of environment. The simplest DDD models use a dielectric constant of εr, which means that electrostatic interactions have an r^{-2} dependence as opposed to r^{-1}. This formulation has served to minimize the contribution of electrostatics to the force field[3] and is computationally efficient. More sophisticated formulations of the DDD model have also been proposed.[50,51] Wisz and Hellinga[52] improved the DDD model by parameterizing dielectric constants and solvation parameters for different regions of the protein. The 24 derived parameters were optimized to reproduce hundreds of experimental pKa values. Their model includes terms for solvation as well as electrostatic interactions and is extremely computationally efficient.

Surface-area-based solvation models have been used in many structure-based energy calculations, including the ORBIT protein design force field.[5,53,54] Although these functions have been successful at modeling non-polar solvation, the basis for using surface area burial or exposure as a measure of solvation energy for polar or charged amino acids is less apparent. Electrostatic interactions act over a longer range than the van der Waals and cavity energies that make up the non-polar solvation energy term.

The Lazardis and Karplus (LK)[55] semi-empirical model measures the desolvation of one atom by another as the product of the volume of the desolvating atom and a solvation free energy density around the desolvated atom. Therefore solute atoms within the first solvation sphere of a particular atom have a larger desolvating effect than atoms further away. The LK model is computationally efficient and can be applied to both polar and nonpolar solvation. The physical assumption that makes the LK model so simple is that atomic desolvation terms are additive. The LK model was re-parameterized by Baker and coworkers and used in the design of a novel fold.[56] The implementation of the LK model into protein design force fields only includes a solvation energy term separate from electrostatic interaction energy. A reported method for using the LK strategy of summing over atomic desolvation terms to approximate effective dielectric constants for electrostatic interactions may also be applicable to protein design energy functions.[50]

Looking ahead

Implementing more sophisticated electrostatic models into highly parameterized protein design force fields brings to the forefront the issue of balancing electrostatics with other protein design force field terms. It has been shown that implementing a series of theoretically improved polar solvation models in a balanced protein design force field does not necessarily lead to improved decoy discrimination.[28] Kuhlman and Baker[6] have reported a readily implemented approach to global optimization of protein design force field, including electrostatics and solvation.

Force field balancing will be futile without considering certain important aspects of protein thermodynamics. Improved models for entropy will be necessary to correctly model salt bridges on the surface of proteins, which may be energetically neutral despite their favorable electrostatic energy. The use of an empirical hydrogen bonding potential[57,58] has been a key component in many successful designs. Continuum models do not capture the covalent character and resulting orientation dependence of hydrogen bonds. As such, an energy function that not only models electrostatics properly but also recognizes favorable hydrogen bonding geometry will be necessary. Morozov et al.[51] reported a balance between hydrogen bonding and a DDD model that discriminated native states from decoys more successfully than either term alone. Currently, energy terms are modeled within the limitation imposed by fixed protein backbone design. This may have a particular impact on electrostatics: if in reality a protein has significant conformational heterogeneity, then the energy of a static salt bridge may not be meaningful.

The combinatorial explosion of protein design has made computationally demanding energy models intractable. If more detailed electrostatic models are to be implemented, there needs to be an effort to develop numerical methods that take advantage of the near redundancy of calculating millions of pairwise rotameric energies. Such methods will be necessary to bring electrostatic modeling up to the challenge of the larger and more ambitious design projects that lie ahead.

Update: May 2008

Since this review article was published in late 2005, several advances have been made in the field of enzyme design. Baker and coworkers have reported the design of enzymes that catalyze the kemp elimination and retro-aldol reactions.[59,60] These enzymes were designed using a protocol in which contacts to a transition state analog are optimized by quantum mechanical calculations. The resulting constellation of catalytic sidechains is then matched with compatible backbone structures, followed by rounds of sequence design around the active site and continuous minimization of the structure. The most active of the designed retro-aldol enzymes use a bound water molecule in the active site. The authors point to the failure of catalytic motifs that use a polar amino acid in the same capacity as evidence for the difficulty in balancing desolvation and polar interactions in catalytic networks.

In addition to the progress in enzyme design, several new methods for modeling polar solvation in protein design energy functions have been reported. McCammon and coworkers developed a method called CIRSE that uses a set of basis functions based on distance-dependent dielectric and solvent exclusion models with parameters trained to reproduce PB/SA calculations.[61] The CIRSE model has been adapted for protein docking and design.[62] Simonson and coworkers have reported a residue pairwise GB model that extends the traditional atom-based Born radii formulation to "residue Born radii" that reflect the burial of each residue.[63] We have reported an update to our two-body PB model that uses generic sidechains to approximate positions with unknown identity.[64] We also showed that this model does better at approximating a full PB calculation than the LK solvent exclusion and the distance-dependent dielectric models. These new

solvation models have yet to be experimentally validated (see Chapter 5 of this thesis). Together with the advances of Baker and coworkers for identifying scaffolds for novel active sites, more accurate modeling of electrostatics should accelerate the pace of computational enzyme design.

Acknowledgements

This work was funded by Howard Hughes Medical Institute, the Ralph M. Parsons Foundation, the Defense Advanced Research Projects Agency, the Institute for Collaborative Biotechnologies (grant DAAD19-03-D-0004 from the U.S. Army Research Office) (SLM), and the National Science Foundation (CLV). We wish to thank Benjamin D. Allen for comments on the manuscript.

References

1. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.

2. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87.

3. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513.

4. Mendes, J., Guerois, R. & Serrano, L. (2002). Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **12**, 441–446.

5. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent–accessible surface areas. *Fold. Des.* **3**, 253–258.

6. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10383–10388.

7. Raha, K., Wollacott, A. M., Italia, M. J. & Desjarlais, J. R. (2000). Prediction of amino acid sequence from structure. *Protein Sci.* **9**, 1106–1119.

8. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149.

9. Warshel, A. (1998). Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *J. Biol. Chem.* **273**, 27035–27038.

10. Bolon, D. N., Voigt, C. A. & Mayo, S. L. (2002). De novo design of biocatalysts. *Curr. Opin. Chem. Biol.* **6**, 125–129.

11. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.

12. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.

13. Lee, L. P. & Tidor, B. (2001). Barstar is electrostatically optimized for tight binding to barnase. *Nat. Struct. Biol.* **8**, 73–76.

14. Sheinerman, F. B. & Honig, B. (2002). On the role of electrostatic interactions in the design of protein-protein interfaces. *J. Mol. Biol.* **318**, 161–177.

15. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein–protein interaction specificity. *Nat. Struct. Biol.* **11**, 371–379.

16. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A Relationship between Protein Stability and Protein Function. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 452–456.

17. Hendsch, Z. S. & Tidor, B. (1994). Do Salt Bridges Stabilize Proteins — a Continuum Electrostatic Analysis. *Protein Sci.* **3**, 211–226.

18. Lumb, K. J. & Kim, P. S. (1995). A Buried Polar Interaction Imparts Structural Uniqueness in a Designed Heterodimeric Coiled–Coil. *Biochem.* **34**, 8642–8648.

19. Lazar, G. A. & Handel, T. M. (1998). Hydrophobic core packing and protein design. *Curr. Opin. Chem. Biol.* **2**, 675–679.

20. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475.

21. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* **305**, 619–631.

22. Korkegian, A., Black, M. E., Baker, D. & Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. *Science* **308**, 857–860.

23. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* **316**, 189–199.

24. McDonald, I. K. & Thornton, J. M. (1994). Satisfying Hydrogen–Bonding Potential in Proteins. *J. Mol. Biol.* **238**, 777–793.

25. Bolon, D. N. & Mayo, S. L. (2001). Polar residues in the protein core of Escherichia coli thioredoxin are important for fold specificity. *Biochem.* **40**, 10047–10053.

26. Bolon, D. N., Marcus, J. S., Ross, S. A. & Mayo, S. L. (2003). Prudent modeling of core polar residues in computational protein design. *J. Mol. Biol.* **329**, 611–622.

27. Desjarlais, J. R. & Clarke, N. D. (1998). Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.* **8**, 471–475.

28. Jaramillo, A. & Wodak, S. J. (2005). Computational protein design is a challenge for implicit solvation models. *Biophys. J.* **88**, 156–171.

29. Pokala, N. & Handel, T. M. (2001). Review: Protein design — where we were, where we are, where we're going. *J. Struct. Biol.* **134**, 269–281.

30. Kraemer–Pecore, C. M., Wollacott, A. M. & Desjarlais, J. R. (2001). Computational protein design. *Curr. Opin. Chem. Biol.* **5**, 690–695.

31. Park, S., Xi, Y. & Saven, J. G. (2004). Advances in computational protein design. *Curr. Opin. Struct. Biol.* **14**, 487–494.

32. Kortemme, T. & Baker, D. (2004). Computational design of protein–protein interactions. *Curr. Opin. Chem. Biol.* **8**, 91–97.

33. Hellinga, H. W. (1996). Metalloprotein design. *Curr. Opin. Biotech.* **7**, 437–441.

34. Hellinga, H. W. (1998). The construction of metal centers in proteins by rational design. *Fold. Des.* **3**, R1–R8.

35. Benson, D. E., Wisz, M. S. & Hellinga, H. W. (1998). The development of new biotechnologies using metalloprotein design. *Curr. Opin. Biotech.* **9**, 370–376.

36. Warshel, A. & Papazyan, A. (1998). Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.* **8**, 211–217.

37. Orozco, M. & Luque, F. J. (2000). Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **100**, 4187–4225.

38. Simonson, T. (2001). Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.* **11**, 243–252.

39. Marshall, S. A., Vizcarra, C. L. & Mayo, S. L. (2005). One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein Sci.* **14**, 1293–1304.

40. Georgescu, R. E., Alexov, E. G. & Gunner, M. R. (2002). Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophys. J.* **83**, 1731–1748.

41. Tanford, C. & Kirkwood, J. G. (1957). Theory of Protein Titration Curves .1. General Equations for Impenettrable Spheres. *J. Am. Chem. Soc.* **79**, 5333–5339.

42. Kirkwood, J. G. (1934). Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions. *J. Chem. Phys.* **2**, 351–361.

43. Havranek, J. J. & Harbury, P. B. (1999). Tanford–Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11145–11150.

44. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**, 45–52.

45. Bashford, D. & Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129–152.

46. Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A. & Brooks, C. L. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **25**, 265–284.

47. Zhu, J., Alexov, E. & Honig, B. (2005). Comparative study of generalized Born models: Born radii and peptide folding. *J. Phys. Chem. B* **109**, 3008–3022.

48. Pokala, N. & Handel, T. M. (2004). Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **13**, 925–936.

49. Dominy, B. N. & Brooks, C. L. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **103**, 3765–3773.

50. Mallik, B., Masunov, A. & Lazaridis, T. (2002). Distance and exposure dependent effective dielectric function. *J. Comput. Chem.* **23**, 1090–1099.

51. Morozov, A. V., Kortemme, T. & Baker, D. (2003). Evaluation of models of electrostatic interactions in proteins. *J. Phys. Chem. B* **107**, 2075–2090.

52. Wisz, M. S. & Hellinga, H. W. (2003). An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins* **51**, 360–377.

53. Eisenberg, D. & Mclachlan, A. D. (1986). Solvation Energy in Protein Folding and Binding. *Nature* **319**, 199–203.

54. Zhang, N. G., Zeng, C. & Wingreen, N. S. (2004). Fast accurate evaluation of protein solvent exposure. *Proteins* **57**, 565–576.

55. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133–152.

56. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.

57. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.

58. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation–dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *J. Mol. Biol.* **326**, 1239–1259.

59. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert,

D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro–aldol enzymes. *Science* **319**, 1387–1391.

60. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., Dechancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195.

61. Cerutti, D. S., Ten Eyck, L. F. & McCammon, J. A. (2005). Rapid estimation of solvation energy for simulations of protein–protein association. *J. Chem. Theor. Comp.* **1**, 143–152.

62. Cerutti, D. S., Jain, T. & McCammon, J. A. (2006). CIRSE: A solvation energy estimator compatible with flexible protein docking and design applications. *Protein Sci.* **15**, 1579–1596.

63. Archontis, G. & Simonson, T. (2005). A residue-pairwise generalized born scheme suitable for protein design calculations. *J. Phys. Chem. B* **109**, 22667–22673.

64. Vizcarra, C. L., Zhang, N., Marshall, S. A., Wingreen, N. S., Zeng, C. & Mayo, S. L. (2008). An improved pairwise decomposable finite–difference Poisson–Boltzmann method for computational protein design. *J. Comput. Chem.* **29**, 1153–1162.

Chapter 3

One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations

The text of this chapter is adapted from a published manuscript that was co-authored with Shannon A. Marshall and Professor Stephen L. Mayo

S.A. Marshall, C.L. Vizcarra, and S.L. Mayo *Protein Science* **14**, 1293–1304 (2005).

Abstract

Successfully modeling electrostatic interactions is one of the key factors required for the computational design of proteins with desired physical, chemical, and biological properties. In this paper, we present formulations of the finite difference Poisson-Boltzmann (FDPB) model that are pairwise decomposable by sidechain. These methods use reduced representations of the protein structure based on the backbone and one or two sidechains in order to approximate the dielectric environment in and around the protein. For the desolvation of polar sidechains, the two-body model has a 0.64 kcal mol^{-1} RMSD compared to FDPB calculations performed using the full representation of the protein structure. Screened Coulombic interaction energies between sidechains are approximated with an RMSD of 0.13 kcal mol^{-1}. The methods presented here are compatible with the computational demands of protein design calculations and produce energies that are very similar to the results of traditional FDPB calculations.

Introduction

Electrostatic interactions are often critical determinants of protein structure and function. In an earlier protein design study, an overly simplistic electrostatic model was found to incorporate destabilizing electrostatic interactions into the designed proteins.[1] Energies calculated using the finite difference Poisson-Boltzmann (FDPB) model,[2–4] a more sophisticated model for describing the electrostatic potential in proteins, correlated more strongly with experimentally determined stability. However, FDPB calculations, as normally performed, are computationally too costly for most protein design calculations.

Computational protein design algorithms[5–8] have relied on simple, often empirical methods to model electrostatic interactions between charged and polar protein groups and the desolvation of polar and charged sidechains. For example, the ORBIT (Optimization of Rotamers by Iterative Techniques) protein design force field uses Coulomb's law with a distance-dependent dielectric and an explicit hydrogen bond term to describe interactions between polar and charged groups and either a penalty for the burial of polar hydrogens or a penalty for the burial of polar surface area.[5,9] Havranek and Harbury have developed a modified Tanford-Kirkwood model to describe electrostatic interactions and applied it to the design of homodimeric and heterodimeric coiled coils.[10,11] Baker and coworkers have used a volume-based solvent exclusion model to describe the desolvation of polar groups,[12] along with a distance-dependent dielectric model, in the successful design of a novel protein fold.[13] Hellinga and coworkers have empirically derived a large number of dielectric constants and interaction parameters to describe polar desolvation as well as charge-charge and charge-polar interactions between protein groups.[14] Finally,

Pokala and Handel have developed a method for calculating Born radii in the context of protein design calculations.[15]

Here, we describe a method for modeling electrostatic interactions in protein design calculations using a limited number of FDPB calculations performed with simplified surface representations. Typically, FDPB calculations require atomic coordinates for the protein backbone and all sidechains in order to define the spatial regions that correspond to the low dielectric protein and high dielectric solvent. In protein design calculations, each possible rotameric sequence (a rotamer is a low energy amino acid sidechain conformation), will have a unique structure and require an independent FDPB calculation. Because the combinatorial complexity of design calculations is often astronomically large, it is not tractable to perform an independent calculation for each possible structure. Instead, we determine the electrostatic energy for each sidechain or pair of sidechains by performing FPDB calculations using simplified structures that include only the backbone and one or two sidechains. The total energy is then obtained by summing the contribution of each sidechain and sidechain pair.

Like the other electrostatic models that have been used for design, the simplified surface approach possesses the computational efficiency required for combinatorially complex protein design calculations. The method is two-body decomposable (meaning that each energy term depends on the identity and conformation of at most two amino acid sidechains) and therefore compatible with deterministic search algorithms such as Dead End Elimination (DEE)[16–18] that are often used for sequence selection. The two-body FDPB method described in this paper allows for calculation of both desolvation energies and electrostatic interactions between polar protein groups using a minimal

number of free parameters. It explicitly captures the impact of sequence changes on the structure of the protein surface, which defines the boundary between the low dielectric protein and the high dielectric solvent. Finally, it efficiently produces energies that correlate well with standard FDPB methods, providing the accuracy demanded by protein design problems.

Strategies for incorporating FDPB methods into protein design calculations

In this study, we have used the FDPB solver from the computer program DelPhi[4] to calculate electrostatic energies for 24 proteins selected from a group of 500 high resolution protein X-ray structures compiled by Richardson and coworkers. The results of these "exact" FDPB calculations were compared to the results of a tractable number of FDPB calculations performed using simplified surface representations that require knowledge of the identity and conformation of no more than two amino acid sidechains at a time in order to assess the accuracy of the simplified surface approximation.

Polar protein groups can form favorable electrostatic interactions with the solvent; we refer to the resulting energies as electrostatic solvation energies. The difference between the electrostatic solvation energy of a polar group in the folded state versus the unfolded state is the desolvation energy. In design calculations, the backbone conformation is typically held fixed. As shown in Figure 3-1A, the desolvation energy of the protein backbone can therefore be defined as the difference between the electrostatic solvation energy of the backbone in the presence of all of the protein's sidechains versus the electrostatic solvation energy of the isolated backbone (a reference state that remains constant in the design calculation). As shown in Figure 3-2A, the desolvation energy of a

sidechain is defined as the difference between the electrostatic solvation energy of the sidechain in the context of the folded protein versus the electrostatic solvation energy of the sidechain and local backbone alone, where the local backbone is defined by the atoms: CA(*i*-1), C(*i*-1), O(*i*-1), N(*i*), H(*i*), CA(*i*), C(*i*), O(*i*), N(*i*+1), H(*i*+1), and CA(*i*+1).

Electrostatic interactions between polar protein groups and the solvent also act to screen Coulombic interactions within a protein. The screening energy is generally opposite in sign and weaker in magnitude than the Coulombic energy for a given interaction. The procedures used to calculate sidechain/backbone and sidechain/sidechain screening energies are shown in Figures 3-2A and 3-3A, respectively. In all cases, the screening energies and Coulombic energies are added to yield "screened Coulombic energies", and the screened Coulombic energies predicted by the different electrostatic models are then compared. As solvation energies are strongly anti-correlated with Coulombic energies, comparison of screened Coulombic energies but not screening energies alone is appropriate for the validation of approximate electrostatic models.[19]

For compatibility with the ORBIT protein design procedure, we have calculated backbone desolvation energies, sidechain desolvation energies, sidechain/backbone interaction energies, and sidechain/sidechain interaction energies separately. The total electrostatic energy of each rotameric state of a protein is then the sum of the backbone desolvation energy (ΔG^{bb}_{desolv}), the desolvation energy of each sidechain *i* (ΔG^{i}_{desolv}), the screened Coulombic interaction between each sidechain *i* and the backbone ($\Delta G^{i/bb}_{screenedCoul}$), and the screened Coulombic interaction between each pair of sidechains *i* and *j* ($\Delta G^{i/j}_{screenedCoul}$):

$$\Delta G^{protein}_{electrostatic} = \Delta G^{bb}_{desolv.} + \sum_i (\Delta G^{i}_{desolv.} + \Delta G^{i/bb}_{screenedCoul.}) + \frac{1}{2}\sum_i \sum_{j \neq i} \Delta G^{i/j}_{screenedCoul.} \quad (1)$$

When calculating the "exact" FDPB energies, each of the above terms is calculated using all of the protein atoms to define the low dielectric protein region versus the high dielectric solvent region.

One-body FDPB decomposition

Several physical properties of proteins can be calculated using information derived from the protein surface. While protein surfaces cannot be perfectly represented using pairwise decomposable methods, earlier protein design studies have demonstrated that pairwise or sequence independent approximations can yield satisfactory results for hydrophobic solvation and binary patterning, respectively.[20,21] Similarly, it may be possible to obtain accurate estimates of the FDPB energies obtained using all the atomic coordinates to define the surface from FDPB energies obtained using simplified models for the protein surface that require knowledge of only one or two sidechain conformations at a time.

Since the protein backbone is fixed during design calculations, an approximate one-body (i.e., one sidechain rotamer) surface can be obtained using the atoms from the protein backbone and the sidechain of interest only. It is necessary to include the sidechain of interest when defining the protein surface to ensure that all protein charges are located in the low dielectric protein region rather than the high dielectric solvent region. The one-body backbone desolvation energy, which is an approximation of the desolvation of the backbone by each sidechain, is calculated as the difference in solvation energy between the one-body folded state (which includes only the sidechain of interest

and the backbone) and the isolated backbone, as shown in Figure 3-1B. The total backbone desolvation energy for each protein is approximated as the sum of the one-body backbone desolvation energies of each of its sidechains. As is shown in Figure 3-2B, one-body sidechain desolvation energies are calculated as the difference in solvation energy between the one-body folded state (which includes only the sidechain of interest and the backbone) and the unfolded state (which includes sidechain *i* and the local backbone). The one-body sidechain/backbone screened Coulombic energy of each sidechain is calculated using the model in Figure 3-2C.

To test the accuracy of the one-body decomposition, we calculated the backbone desolvation energies, sidechain desolvation energies, and sidechain/backbone screened Coulombic energies for the set of 24 proteins. Backbone desolvation energies can be calculated reasonably well by summing the desolvation induced by the presence of each sidechain, as shown in Figure 3-4A. Using the one-body decomposition, the backbone desolvation energy resulting from each sidechain can be considered as a component of the sidechain/backbone energy of the sidechain in design calculations. The extent to which backbone desolvation energy depends on protein sequence and sidechain conformations is not yet fully understood. Avbelj, Baldwin, and coworkers, however, have reported the importance of backbone desolvation in determining amino acid secondary structure propensities. [22–24]

The one-body approximation grossly underestimates the majority of the sidechain desolvation and sidechain/backbone screened Coulombic energies, as shown in Figures 3-4B and 3-4C, respectively. The one-body model neglects the contribution of the other sidechains to the dielectric environment of the sidechain of interest, resulting in an

excessively solvated folded state. Deviations between the one-body and exact FDPB results are especially pronounced for large magnitude desolvation and screened Coulombic energies, which tend to occur in environments with a low effective dielectric.

Two-body FDPB decomposition

More accurate energies can be obtained using two-body methods (i.e., methods including two sidechain rotamers), in which the total sidechain desolvation or sidechain/backbone screened Coulombic energy for each sidechain *i* is defined as the sum of its one-body energy and the two-body perturbation energies for each other sidechain *j*. As shown in Figures 3-2B and 3-2C, the perturbation energy of each other sidechain is defined as the difference between the two-body energy, which is calculated using the backbone and two sidechains to define the dielectric boundary, and the one-body energy calculated previously.

Incorporating the effects of other sidechains using the two-body perturbation method allows accurate calculation of electrostatic energies, as shown in Table 3-1 and Figures 3-5A and 3-5B. Five outlier points, representing five different amino acid types from four structures, were observed to have large errors in their two-body sidechain desolvation energies, as shown in Figure 3-5A. These outliers likely arise from grid placement artifacts, a source of error in FDPB calculations that has been described previously.[2] Accurate two-body desolvation energies can be obtained for these five points by slightly altering the position of the molecule relative to the grid (data not shown).

The two-body approximation systematically underestimates the magnitude of the sidechain desolvation energy. The systematic error in the two-body desolvation energy

was minimized by linearly scaling the two-body perturbation energy. The set of 24 structures was divided into two sets of 12 structures, and a scaling parameter, α, was derived by a linear least-squares fit for each set (with the five outlier points removed). The robustness of the scaling parameter was tested by cross-validation, as shown in Table 3-2, and sensitivity analysis, as shown in Figure 3-6A. The error in the two-body sidechain desolvation is reasonably insensitive to α around the optimal α value, and both sets have similar dependence on α, suggesting that this scaling parameter should be used in routine calculations.

In the one-body FDPB method, we calculated sidechain and backbone desolvation energies and sidechain/backbone screening energies, but not sidechain/sidechain screening energies. Simply multiplying the one-body potential generated by sidechain *i* by the partial atomic charges of sidechain *j* is not very accurate (data not shown), especially for charged atoms located at or beyond the dielectric boundary defined by sidechain *i* and the protein backbone. Sidechain/sidechain screened Coulombic energies were calculated using a two-body decomposable method that uses only the backbone and two sidechains of interest to define the dielectric boundary, as shown in Figure 3-3B. Although the two-body model systematically over-screens the Coulombic interactions, the accuracy obtained using a two-body FDPB decomposition is quite good, as shown in Table 3-1 and Figure 3-5C. The two-body approximation is probably less accurate for certain large interaction energies due to increased sensitivity to the shape of the dielectric boundary in regions of large electrostatic potential.

Analysis of the sidechain desolvation and sidechain/backbone screened Coulombic energies indicates that, in most cases, the perturbation caused by a second

sidechain is negligible. The small fraction of two-body perturbations that contribute significantly to the desolvation or sidechain/backbone energies involve pairs of residues that are close in space. Furthermore, sidechain/sidechain interaction energies for residues that are not close in space are typically small in magnitude and may be approximated using a simpler electrostatic model. We performed additional calculations in which two-body perturbations were calculated only for pairs that separated by less than 6 Å or 4 Å. As shown in Table 3-1, we observe a slight decrease in accuracy as the distance cutoff is decreased from infinity to 6 Å to 4 Å. This arises from an increased underestimation of the sidechain desolvation energies and sidechain/backbone screened Coulombic energies, as well as increased inaccuracy in defining the dielectric environment, as fewer pairs are included.

When calculating screened Coulombic energies, the interaction of sidechain pairs separated by more than a distance cutoff of 6 Å or 4 Å was approximated by a distance-dependent Coulombic model and the two-body FDPB model was applied only to pairs that are close in space. The two sets of protein structures used for the α parameterization were used to derive the optimal distance dependent dielectric values for pairs separated by distances greater than the cutoff. The dielectrics derived for each set are similar, and the errors in the two-body approximation with the cutoffs are comparable to the error in the full two-body calculation including all pairs, as shown in Table 3-3. The sensitivity of the error and correlation with the exact FDPB energies to the dielectric value is shown in Figures 3-6B and 3-6C.

Considering only a limited subset of pairs significantly reduces the total calculation time, which is crucial since the number of pairs in a design calculation is

often large. For instance, the reported surface design calculation for engrailed homeodomain considers 15,000,000 rotamer pairs.[1] The FDPB calculation for this number of pairs would require approximately three weeks of CPU time on a cluster of 128 IBM PowerPC 970 processors running at 1.6 GHz. The time required to complete the two-body calculation can be reduced to less than one day of CPU time by applying a distance cutoff of 4.0 Å.

It has been shown that, for a series of designed homeodomain variants, there is a correlation between experimental stability and exact FDPB electrostatic energies plus ORBIT van der Waals energies.[1] In order to assess the predictive power of the two-body method presented here, we have compared the two-body FDPB energies to these experimental results. For each variant, the sum of all two-body sidechain/backbone and sidechain/sidechain screened Coulombic energies and the sum of all two-body sidechain desolvation energies were added to the ORBIT van der Waals energies. As shown in Figure 3-7, the two-body FDPB energies are able to predict, with accuracy close to that of the exact FDPB calculations, trends in experimental stabilities of six of the seven variants tested, including the wild-type protein and NC3-Ncap, the most stable variant.

Additional considerations

Thus far, we have developed and tested new electrostatic models for protein design calculations by maximizing the agreement between the approximate desolvation and screened Coulombic energies with the exact FDPB energies. While even "exact" FDPB energies are an approximation of the true electrostatic energy of the system, it is probable that, in the context of design calculations, the accuracy of the structural model

will be a greater source of error than the limitations of the underlying FDPB model. To maximize computational efficiency, most protein design methods use a fixed backbone, discrete sidechain rotamers, and a very simple model of the unfolded state. As a result, certain errors in electrostatic energies can be observed in design calculations. For example, the energetic benefit of surface salt bridges is overestimated if the entropic cost of locking flexible sidechains into a single conformation is not considered. Similarly, the folded state stability conferred by interactions that are populated in the unfolded state, such as i, $i\pm2$ sidechain/backbone interactions, is overestimated if the unfolded state is modeled as the sidechain and local backbone only.

Based on a single study of electrostatics in designed proteins,[1] either exact or two-body FDPB energies (with large magnitude sidechain-sidechain interactions truncated) are sufficiently accurate to provide a reasonable correlation with experimentally determined stability, as shown in Figure 3-7. Additional experimental studies will be required to assess the performance of the two-body decomposable model in the design of proteins with specific catalytic or binding properties. In cases where accurate modeling of electrostatics is especially critical, more sophisticated structural models, such as the flexible rotamer model[25] and explicit modeling of alternate backbone conformations,[13] may prove useful.

Conclusions

Accurate electrostatic models, including the FDPB model, require knowledge of the full tertiary structure of the protein. As a result, these models cannot be applied directly to protein design calculations, which often consider over 10^{50} possible protein structures. While it is not possible to explicitly calculate electrostatic energies in each structural environment, it is also not prudent to neglect changes in the shape of a protein's surface that result from modifying the protein sequence.

We have found that it is possible to obtain accurate electrostatic energies using simplified surface models that depend on the identity and conformation of the protein backbone and only one or two sidechains at a time. The success of the two-body FDPB method suggests that it is critical to define the surface accurately in the immediate vicinity of the partial charges that are "generating" and "feeling" the electrostatic potential in each calculation. The results also suggest that it is important to account for desolvation and screening due to other nearby sidechains, but that the effects of each other sidechain are fairly independent and can be captured pairwise. Finally, we have found that the effects of sequence-dependent variation in the dielectric boundary can be neglected if the perturbations are reasonably far removed from the partial charges that are "generating" or "feeling" the electrostatic potential in a given calculation.

Efficient and accurate electrostatic models are also critical for protein folding and docking calculations. The simplified surface methods discussed here could be used to explore different sidechain orientations given a fixed backbone conformation. Similarly, derivatives of a small molecule scaffold, such as those generated by combinatorial chemistry methods, could be modeled. However, folding and docking calculations

typically sample a large number of backbone conformations or relative molecular orientations. Since each backbone conformation would require an independent set of one- or two-body FDPB calculations, the computational demands of folding and docking calculations would be far greater than those for design.

The stability of designed proteins has already been demonstrated to be sensitive to the quality of the electrostatic model used in the design calculations. It is likely that electrostatic interactions are at least as important in determining the functional properties of proteins, including binding and catalysis. As a result, the development and testing of accurate electrostatic models is likely to significantly aid in the design of proteins with desired physical, chemical, and biological properties.

Materials and Methods

Test set of proteins. All calculations were performed using proteins selected from a group of 500 high-resolution protein X-ray structures, including computationally optimized hydrogen atom locations, compiled by Richardson and coworkers (http://kinemage.biochem.duke.edu/databases/top500.php). Structural coordinates were derived from PDB entries 1IGD, 1MSI, 1KP6, 1OPD, 1FNA, 1MOL, 2ACY, 1ERV, 1DHN, 1WHI, 3CHY, 1ELK, 2RN2, 1HKA, 3LZM, 1AMM, 1XNB, 153L, 1BK7, 2PTH, 1THV, 1BS9, 1AGJ, and 2BAA, corresponding to the β1 domain of Streptococcal protein G, type III antifreeze protein, alpha subunit of killer toxin KP6, S46A mutant of *E. coli* phosphotransferase, fibronection cell-adhesion module type III, monellin, bovine acyl-phosphatase, C73S mutant of human thioredoxin, 7,8-dihydroneopterin aldolase, L14 ribosomal protein, CheY, VHS domain of TOM1, ribonuclease H, pyrophosphokinase, T4 lysozyme, gamma-B-crystallin, xylanase, goose lysozyme, ribonuclease MC1, peptidyl-tRNA hydrolase, thaumatin, acetylxylan esterase, epidermolytic toxin A from *S. aureus*, and endochitinase, respectively. Only the "A" chain was used for monellin, the VHS domain of TOM1, and epidermolytic toxin A.

Exact FDPB calculations. Finite difference solutions to the linearized Poisson-Boltzmann equation were obtained using the FDPB solver from the computer program DelPhi[4] with a grid spacing of 2.0 grids $Å^{-1}$, an interior dielectric of 4.0, an exterior dielectric of 80.0, a salt concentration of 0.050 M, and a probe radius of 1.4 Å. The grid size was selected for each protein so that its backbone atoms fill 70% of the grid. The coordinates of each protein were mapped onto the grid in exactly the same way in all of the calculations to minimize errors due to changing grid placement. The PARSE parameter set charges and atomic radii[26] were used in all FDPB calculations. Proline residues and cysteine residues in disulfide bonds were considered part of the backbone in

all calculations. All Arg and Lys residues were modeled with a +1 net charge and all Asp and Glu residues were modeled with a -1 charge. All FDPB energies were converted to units of kcal mol^{-1} using the relation kT = 0.593 kcal mol^{-1} at 25 °C.

In the FDPB model, electrostatic solvation energies are obtained by multiplying the appropriate atomic charges, q, by the reaction field potential, ϕ, at the location of each charge. In the following equations, the reaction field potential, ϕ, is labeled with a superscript that indicates which atoms were used to define the dielectric boundary and with a subscript that indicates which atoms were assigned non-zero partial atomic charges when calculating the reaction field potential. The entire protein is referred to as "*all*", the protein backbone is referred to as "*bb*", individual protein side chains are referred to as "*i*" or "*j*", and a sidechain with its local backbone is referred to as "*ib*". A factor of 1/2 appears in the desolvation energy equations to account for the work of solvent polarization in response to the charges on sidechain *i*.

The exact desolvation energy of the backbone ("*bb*"), shown in Figure 3-1A, is defined as the difference between the electrostatic solvation energy of the backbone in the presence of all the protein sidechains and the electrostatic solvation energy of the backbone alone:

$$\Delta G^{bb}_{desolv.} = \frac{1}{2}\sum_{t} q_t(\phi^{all}_{bb} - \phi^{bb}_{bb}) \qquad (2)$$

where each t is a backbone atom, q_t is the partial atomic charge of backbone atom t, ϕ^{all}_{bb} is the reaction field potential at t generated by the set of partial atomic charges on the backbone when all of the protein atoms are used to define the dielectric boundary, and ϕ^{bb}_{bb} is the reaction field potential at t generated by the set partial atomic charges on the backbone when the backbone atoms only are used to define the dielectric boundary.

The exact desolvation energy of a sidechain i, shown in Figure 3-2A, is defined as the difference between the electrostatic solvation energy of the sidechain in the folded state versus the unfolded state:

$$\Delta G^{i}_{desolv.} = \frac{1}{2}\sum_{u} q_u(\phi^{all}_{i} - \phi^{ib}_{i}) \tag{3}$$

where each *u* is an atom in sidechain *i*, q_u is the partial atomic charge of sidechain atom *u*, ϕ^{all}_{i} is the reaction field potential at *u* generated by the set of partial atomic charges on sidechain *i* when all of the protein atoms are used to define the dielectric boundary, and ϕ^{ib}_{i} is the reaction field potential at *u* generated by the set of partial atomic charges on sidechain *i* when the atoms on sidechain *i* and its local backbone are used to define the dielectric boundary. The molecular surface for the sidechain unfolded state model was generated using the sidechain and local backbone and was mapped to the grid exactly as in the folded state calculations. The local backbone was defined to include the following atoms: CA(*i*-1), C(*i*-1), O(*i*-1), N(*i*), H(*i*), CA(*i*), C(*i*), O(*i*), N(*i*+1), H(*i*+1), and CA(*i*+1).

Exact folded state sidechain/backbone screening energies, shown in Figure 3-2A, were obtained using the following equation:

$$\Delta G^{i/bb}_{screening} = \sum_{t} q_t \phi^{all}_{i} \tag{4}$$

where *i* is the sidechain of interest, each *t* is an atom in the backbone, q_t is the partial atomic charge of atom *t*, and ϕ^{all}_{i} is the reaction field potential at *t* generated by the set of partial atomic charges on sidechain *i* when all of the protein atoms are used to define the dielectric boundary. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G^{i/bb}_{screenedCoulombic} = \Delta G^{i/bb}_{screening} + \Delta G^{i/bb}_{Coulombic} \tag{5}$$

where the Coulombic energy is calculated using Coulomb's law with a dielectric constant equal to the dielectric of the protein interior.

Exact sidechain/sidechain interactions, shown in Figure 3-3A, were obtained using a similar method:

$$\Delta G^{i/j}_{screening} = \sum_{v} q_v \phi^{all}_i \tag{6}$$

where *i* and *j* are the sidechains of interest, each *v* is an atom in sidechain *j*, q_v is the partial atomic charge of atom *v*, and ϕ^{all}_i is the reaction field potential at *v* generated by the set of partial atomic charges on sidechain *i* when all of the protein atoms are used to define the dielectric boundary. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G^{i/j}_{screenedCoulombic} = \Delta G^{i/j}_{screening} + \Delta G^{i/j}_{Coulombic} \tag{7}$$

Sidechain/backbone and sidechain/sidechain interaction energies are assumed to be zero in the unfolded state.

One-body FDPB calculations. One-body FDPB energies were calculated for backbone desolvation energies, sidechain desolvation energies, and sidechain/backbone screened Coulombic energies. For each sidechain in the test set, two FDPB calculations are carried out: one with non-zero partial atomic charges assigned to the sidechain and one with non-zero partial atomic charges assigned to the backbone. Folded state solvation energies for the protein backbone were calculated as in the exact FDPB calculations, except that sidechains other than the sidechain of interest were not included:

$$\Delta G^{bb,1-body}_{desolv.} = \frac{1}{2} \sum_{t} q_t (\phi^{i,bb}_{bb} - \phi^{bb}_{bb}) \tag{8}$$

where each *t* is a backbone atom, q_t is the partial atomic charge of backbone atom *t*, $\phi^{i,bb}_{bb}$ is the reaction field potential at *t* generated by the set of partial atomic charges on the backbone when sidechain *i* and the backbone atoms only are used to define the dielectric boundary, and ϕ^{bb}_{bb} is the reaction field potential at *t* generated by the set of partial atomic charges on the backbone when the backbone atoms only are used to define the dielectric boundary, as shown in Figure 3-1B. The total backbone desolvation energy for each

protein is approximated by the sum of the one-body backbone desolvation energies, given by Equation 8, for each of its sidechains.

Sidechain desolvation energies were calculated as in the exact FDPB calculations, except only the sidechain of interest and the backbone were used to construct the folded state dielectric boundary:

$$\Delta G_{desolv.}^{i,1-body} = \frac{1}{2}\sum_{u} q_u(\phi_i^{i,bb} - \phi_i^{ib}) \tag{9}$$

where *i* is the sidechain of interest, each *u* is an atom in sidechain *i*, q_u is the partial atomic charge of atom *u*, $\phi^{i,bb}{}_i$ is the reaction field potential at *u* generated by the set of partial atomic charges on sidechain *i* when sidechain *i* and the backbone atoms only are used to define the dielectric boundary, and $\phi^{ib}{}_i$ is the reaction field potential at *u* generated by the set of partial atomic charges on sidechain *i* when the atoms in sidechain *i* and its local backbone are used to define the dielectric boundary, as shown in Figure 3-2B.

Similarly, sidechain/backbone screened Coulombic energies were calculated as in the exact FDPB calculations, except only the sidechain of interest and the backbone were used to construct the dielectric boundary:

$$\Delta G_{screening}^{i/bb,1-body} = \sum_{t} q_t \phi_i^{i,bb} \tag{10}$$

where *i* is the sidechain of interest, each *t* is a backbone atom, q_t is the partial atomic charge of atom *t*, and $\phi^{i,bb}{}_i$ is the reaction field potential at *t* generated by the set of partial atomic charges on sidechain *i* when sidechain *i* and the backbone atoms only are used to define the dielectric boundary, as shown in Figure 3-2C. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G_{screenedCoulombic}^{i/bb,1-body} = \Delta G_{screening}^{i/bb,1-body} + \Delta G_{Coulombic}^{i/bb} \tag{11}$$

where the Coulombic energy is calculated using Coulomb's law with a dielectric constant equal to the dielectric of the protein interior.

Two-body FDPB calculations. Two-body FDPB sidechain desolvation energies, sidechain/backbone screened Coulombic energies, and sidechain/sidechain screened Coulombic energies were calculated as follows. First, the one-body energies were calculated as described above. Next, two-body perturbation energies were calculated using the atoms in the backbone, *bb*, the sidechain of interest, *i*, and one "perturbing" sidechain, *j*, to define the dielectric boundary. Two-body perturbation energies were calculated using each residue other than the sidechain of interest as the perturbing residue. Total energies were calculated by adding the one-body energy to the sum of the two-body perturbation energies. For each pair of sidechains, two FDPB calculations are carried out, one with non-zero partial atomic charges assigned to each sidechain.

Two-body sidechain desolvation energies were calculated as the sum of a one-body energy and two-body perturbation energies:

$$\Delta G_{desolv.}^{i,2-body} = \Delta G_{desolv.}^{i,1-body} + \sum_{j \neq i} \left[\frac{1}{2} \sum_{u} q_u (\phi_i^{i,j,bb} - \phi_i^{ib}) - \Delta G_{desolv.}^{i,1-body} \right] \qquad (12)$$

where *i* is the sidechain of interest, each *u* is an atom in sidechain *i*, q_u is the partial atomic charge of *u*, and $\phi^{ij,bb}{}_i$ is the reaction field potential at *u* generated by the set of partial atomic charges on side chain *i* when the backbone and side chains *i* and *j* are used to define the dielectric boundary, as shown in Figure 3-2B.

In order to improve the accuracy of the two-body sidechain desolvation energy, a scaling parameter, α, was multiplied by the term in Equation 12 that sums over sidechains *j*. This parameter was fit using two distinct sets of structures. Structure set 1 contained 1IGD, 1KP6, 1FNA, 2ACY, 1DHN, 3CHY, 2RN2, 3LZM, 1XNB, 1BK7, 1THV, and 1AGJ. Structure set 2 contained 1MSI, 1OPD, 1MOL, 1ERV, 1WHI, 1ELK, 1HKA, 1AMM, 153L, 2PTH, 1BS9, and 2BAA. Optimum values of α were determined for each set by linear least squares fit, and a sensitivity analysis was performed by testing values of α between 1.0 and 2.0 at intervals of 0.05.

Two-body sidechain/backbone screened Coulombic energies were calculated as the sum of a one-body energy and two-body perturbation energies:

$$\Delta G^{i/bb,2-body}_{screening} = \Delta G^{i/bb,1-body}_{screening} + \sum_{j \neq i} \left[\sum_{t} q_t \phi^{i,j,bb}_{i} - \Delta G^{i/bb,1-body}_{screening} \right] \quad (13)$$

where i is the sidechain of interest, each t is a backbone atom, q_t is the partial atomic charge of t, and $\phi^{i,j,bb}{}_{i}$ is the reaction field potential at t generated by the set of partial atomic charges on side chain i when the backbone and side chains i and j are used to define the dielectric boundary, as shown in Figure 3-2C. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G^{i/bb,2-body}_{screenedCoulombic} = \Delta G^{i/bb,2-body}_{screening} + \Delta G^{i/bb}_{Coulombic} \quad (14)$$

where the Coulombic energy is calculated using Coulomb's law with a dielectric constant equal to the dielectric of the protein interior.

Two-body sidechain/sidechain calculations were calculated using the same method that was used to calculate the exact sidechain/sidechain screening energies, except that the dielectric boundary is defined using only the backbone and the two sidechains of interest:

$$\Delta G^{i/j,2-body}_{screening} = \sum_{v} q_v \phi^{i,j,bb}_{i} \quad (15)$$

where i and j are the two sidechains of interest, each v is an atom in sidechain j, q_v is the partial atomic charge of atom v, and $\phi^{i,j,bb}{}_{i}$ is the reaction field potential at v generated by the set of partial atomic charges on sidechain i when the backbone and sidechains i and j are used to define the dielectric boundary, as shown in Figure 3-3B. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G^{i/j,2-body}_{screenedCoulombic} = \Delta G^{i/j,2-body}_{screening} + \Delta G^{i/j}_{Coulombic} \quad (16)$$

where the Coulombic energy is calculated using Coulomb's law with a dielectric constant equal to the dielectric of the protein interior. Figure 3-8 outlines how the one- and two-body FDPB calculations described here can be implemented in a protein design code.

For the two-body sidechain desolvation and sidechain/backbone screened Coulombic energy calculations using only pairs that are close in space, the distance between sidechains i and j was defined as the minimum distance between any atom with non-zero partial atomic charge on sidechain *i* and any atom on sidechain *j*. For two-body sidechain/sidechain screened Coulombic energy calculations using only pairs that are close in space, the distance between sidechains *i* and *j* was defined as the minimum distance between any atom with non-zero partial atomic charge on sidechain *i* and any atom with non-zero partial atomic charge on sidechain *j*. In sidechain/sidechain calculations, Coulomb's law was used to calculate the energy of pairs that were farther apart than the cutoff distance. For cutoff distances of both 6.0 Å and 4.0 Å, optimal distance dependent dielectric values were derived by linear least-squares to maximize agreement with the exact FDPB sidechain/sidechain screened Coulombic energies. These dielectric values were tested by cross-validation, and the sensitivity of the error in the two-body approximation with a cutoff was tested by varying the dielectric values.

Two-body energies were calculated for a series of homeodomain variants reported by Marshall et al. (2002). For each variant, FDPB two-body sidechain desolvation energies, two-body sidechain/backbone screened Coulombic energies, and two-body sidechain/sidechain screened Coulombic energies were added to the total ORBIT van der Waals energy. A threshold of ± 0.90 kcal/mol was applied to the sidechain/backbone and

sidechain/sidechain screened Coulombic energies. FDPB calculations were run using parameters described previously.[1]

Acknowledgements

We would like to thank Barry Honig and Emil Alexov for helpful conversations. This work was supported by the Howard Hughes Medical Institute, the Ralph M. Parsons Foundation, an IBM Shared University Research Grant, DARPA, ARO/ICB (S. L. M.), a National Science Foundation graduate research fellowship (C. L. V.), a National Institutes of Health training grant, and the Caltech Initiative in Computational Molecular Biology program, awarded by the Burroughs Wellcome Fund (S. A. M.).

References

1. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* **316**, 189–199.

2. Gilson, M. K., Sharp, K. & Honig, B. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.* **9**, 327–335.

3. Honig, B. & Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149.

4. Rocchia, W., Alexov, E. & Honig, B. (2001). Extending the applicability of the nonlinear Poisson–Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **105**, 6507–6514.

5. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513.

6. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Struct. Fold. Des.* **7**, R105–R109.

7. Kraemer-Pecore, C. M., Wollacott, A. M. & Desjarlais, J. R. (2001). Computational protein design. *Curr. Opin. Chem. Biol.* **5**, 690–695.

8. Mendes, J., Guerois, R. & Serrano, L. (2002). Energy estimation in protein design. *Curr. Opin. Struct. Biol.* **12**, 441–446.

9. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.

10. Havranek, J. J. & Harbury, P. B. (1999). Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11145–11150.

11. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**, 45–52.

12. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133–152.

13. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.

14. Wisz, M. S. & Hellinga, H. W. (2003). An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins* **51**, 360–377.

15. Pokala, N. & Handel, T. M. (2004). Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Sci*. **13**, 925–936.

16. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542.

17. Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side–chains and related spin-glasses. *Biophys. J*. **66**, 1335–1340.

18. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J. Comput. Chem*. **24**, 232–243.

19. Scarsi, M. & Caflisch, A. (1999). Comment on the validation of continuum electrostatics models. *J. Comput. Chem*. **20**, 1533–1536.

20. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent–accessible surface areas. *Fold. Des*. **3**, 253–258.

21. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol*. **305**, 619–631.

22. Avbelj, F. & Moult, J. (1995). Role of electrostatic screening in determining protein main-chain conformational preferences. *Biochem*. **34**, 755–764.

23. Avbelj, F. & Fele, L. (1998). Role of main-chain electrostatics, hydrophobic effect and side–chain conformational entropy in determining the secondary structure of proteins. *J. Mol. Biol*. **279**, 665–684.

24. Avbelj, F., Luo, P. Z. & Baldwin, R. L. (2000). Energetics of the interaction between water and the helical peptide group and its role in determining helix propensities. *Proc. Natl. Acad. Sci. U. S. A*. **97**, 10786–10791.

25. Mendes, J., Baptista, A. M., Carrondo, M. A. & Soares, C. M. (1999). Improved modeling of side–chains in proteins with rotamer–based methods: A flexible rotamer model. *Proteins* **37**, 530–543.

26. Sitkoff, D., Sharp, K. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem*. **98**, 1978–1988.

Table 3-1: Accuracy of the electrostatic models

	RMSD (kcal mol^{-1})	R
A. Backbone desolvation energy		
exact FDPB	-	-
one-body	3.96	0.997
B. Sidechain desolvation energy		
exact FDPB	-	-
one-body	1.93	0.718
two-body[a], all pairs	0.64	0.962
two-body[a], pairs < 6 Å	0.67	0.968
two-body[a], pairs < 4 Å	0.82	0.952
C. Sidechain/backbone screened Coulombic energy		
exact FDPB	-	-
one-body	0.90	0.957
two-body, all pairs	0.36	0.987
two-body, pairs < 6 Å	0.41	0.984
two-body, pairs < 4 Å	0.51	0.979
D. Sidechain/sidechain screened Coulombic energy		
exact FDPB	-	-
two-body, all pairs	0.13	0.948

[a] Statistics were obtained using all data points, including outliers, and without application of α, the scaling parameter for two-body sidechain desolvation.

Table 3-2: Cross-validation of α, the scaling parameter for two-body sidechain desolvation

	RMSD (kcal mol^{-1})	R
Structure set 1		
$\alpha = 1$	0.56	0.967
$\alpha = 1.26$[a]	0.43	0.972
$\alpha = 1.30$[b]	0.43	0.973
Structure set 2		
$\alpha = 1$	0.68	0.971
$\alpha = 1.26$[a]	0.50	0.974
$\alpha = 1.30$[b]	0.50	0.974

[a]The optimal value of α determined using structure set 1
[b]The optimal value of α determined using structure set 2

Table 3-3: Cross-validation of distance-dependent dielectrics for limited pair two-body sidechain/sidechain screened Coulombic interactions

	RMSD[a] (kcal mol^{-1})	R[a]
Structure set 1		
all pairs	0.10	0.968
pairs > 6 Å, $\varepsilon = 5.11r$[b]	0.10	0.960
pairs > 6 Å, $\varepsilon = 4.75r$[c]	0.10	0.957
pairs > 4 Å, $\varepsilon = 5.90r$[d]	0.10	0.955
pairs > 4 Å, $\varepsilon = 5.21r$[e]	0.10	0.947
Structure set 2		
all pairs	0.16	0.934
pairs > 6 Å, $\varepsilon = 5.11r$[b]	0.16	0.926
pairs > 6 Å, $\varepsilon = 4.75r$[c]	0.16	0.923
pairs > 4 Å, $\varepsilon = 5.90r$[d]	0.16	0.924
pairs > 4 Å, $\varepsilon = 5.21r$[e]	0.16	0.917

[a]RMSD and R values are for all pairs in each structure set.
[b]The optimal distance dependent dielectric for pairs separated by > 6 Å in structure set 1
[c]The optimal distance dependent dielectric for pairs separated by > 6 Å in structure set 2
[d]The optimal distance dependent dielectric for pairs separated by > 4 Å in structure set 1
[e]The optimal distance dependent dielectric for pairs separated by > 4 Å in structure set 2

Figure 3-1. Free energy cycles used to calculate (A) exact versus (B) one-body backbone desolvation energies (as shown in Equations 2 and 8, respectively). In each method, the electrostatic potential generated by the backbone is calculated. The key distinctions between the two methods are as follows: the exact calculation uses the protein backbone and all of the sidechains in the protein to define the dielectric boundary, while in the one-body method, the dielectric boundary is defined by the backbone and a single sidechain only. The total one-body desolvation is calculated by summing the desolvation by each sidechain. The parameters used in each FDPB calculation are indicated as follows: the protein backbone, shown in red, was assigned partial atomic charges from the PARSE charge set; the sidechains, shown in gray, were assigned partial atomic charges of 0; the areas drawn in white were assigned a dielectric constant of 4 (protein interior); and the blue areas were assigned a dielectric constant of 80 (water) and a salt concentration of 50 mM.

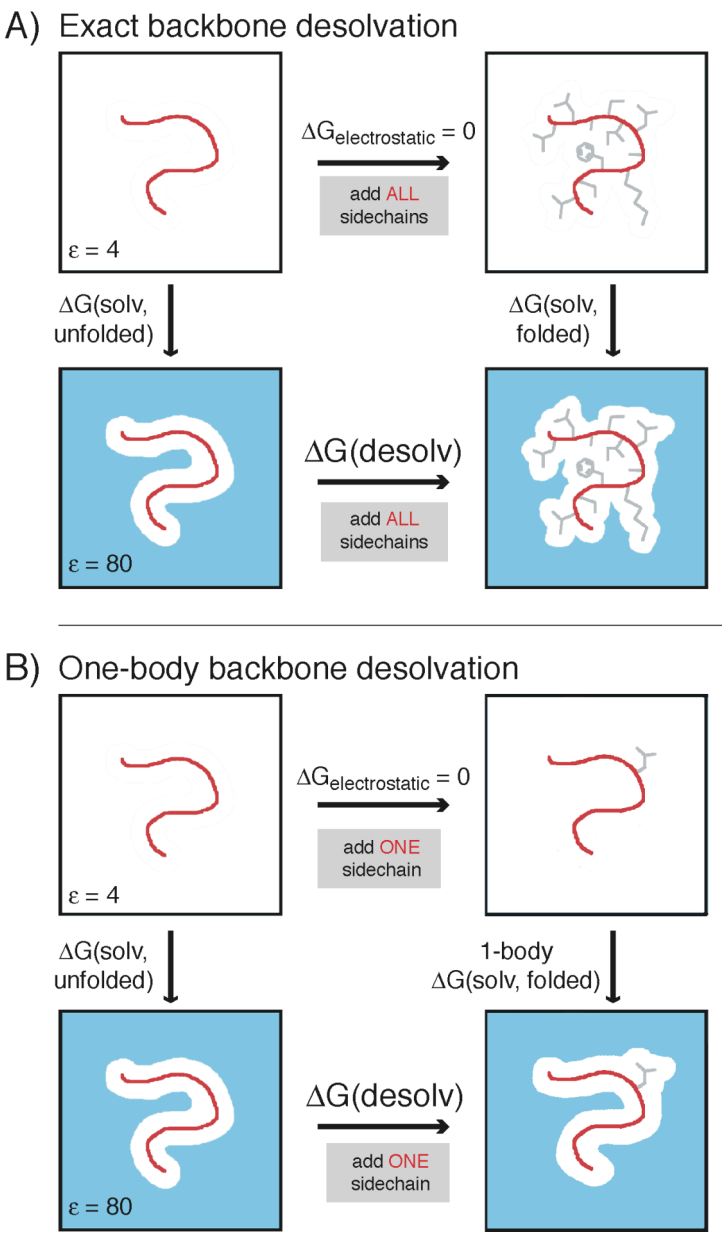

A) Exact backbone desolvation
ΔG electrostatic = 0
add ALL sidechains
ε = 4
ΔG(solv, unfolded)
ΔG(solv, folded)
ΔG(desolv)
add ALL sidechains
ε = 80
B) One-body backbone desolvation
ΔG electrostatic = 0
add ONE sidechain
ε = 4
ΔG(solv, unfolded)
1-body ΔG(solv, folded)
ΔG(desolv)
add ONE sidechain
ε = 80

Figure 3-2. Free energy cycles used to calculate (A) exact sidechain desolvation energies (as shown in Equation 3) and sidechain/backbone screened Coulombic energies (as shown in Equations 4–5) versus one-body and two-body (B) sidechain desolvation energies (as shown in Equations 9 and 12, respectively) and (C) sidechain/backbone screened Coulombic energies (as shown in Equations 10–11 and 13–14, respectively). In each method, the electrostatic potential generated by sidechain *i* is calculated. This potential is multiplied by the charges of sidechain *i* to calculate the solvation energy of *i* and is multiplied by the charges in the backbone to determine the sidechain/backbone screening energy. The key distinctions between the exact, one-body, and two-body methods are as follows. The exact calculation uses the protein backbone and all of the sidechains in the protein to define the dielectric boundary, and a single calculation is used to determine the folded state solvation energy. In the one-body method, the dielectric boundary is defined by the backbone and a single sidechain only. The one-body desolvation energy consists of the desolvation of sidechain *i* by the backbone. In the two-body method, a one-body calculation is first performed as shown in parts (B) and (C), and then the perturbation in the sidechain desolvation energy and the sidechain/backbone screened Coulombic energy that results from adding a second sidechain, *j*, to the low dielectric protein region is determined. The perturbation due to each other sidechain is added to the one-body energy to produce the two-body energy. The parameters used in each FDPB calculation are indicated as follows: sidechain *i*, shown in red, was assigned partial atomic charges from the PARSE charge set; the rest of the protein, when shown in gray, was assigned partial atomic charges of 0; the protein backbone, when shown in green, was assigned partial atomic charges of 0 in the FDPB calculation, but its PARSE partial atomic charges were used to obtain screening energies; the areas drawn in white were assigned a dielectric constant of 4 (protein interior); and the blue areas were assigned a dielectric constant of 80 (water) and a salt concentration of 50 mM.

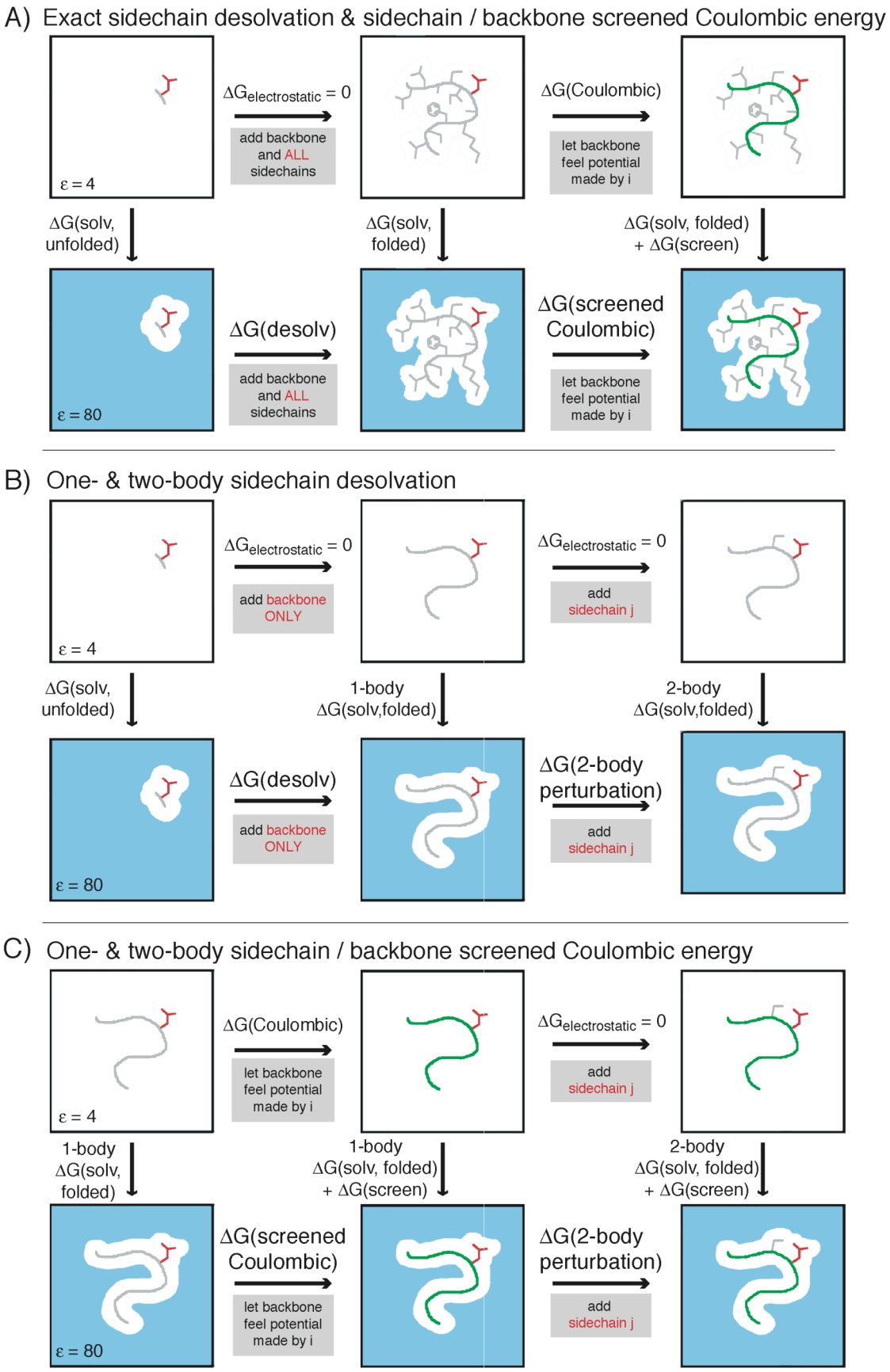
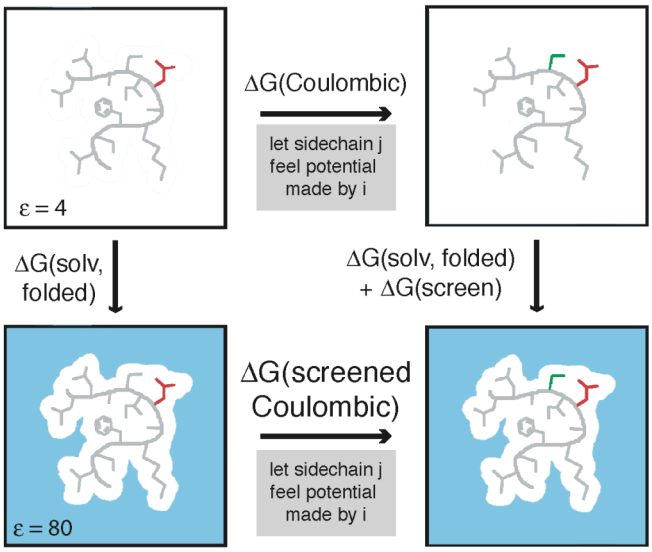
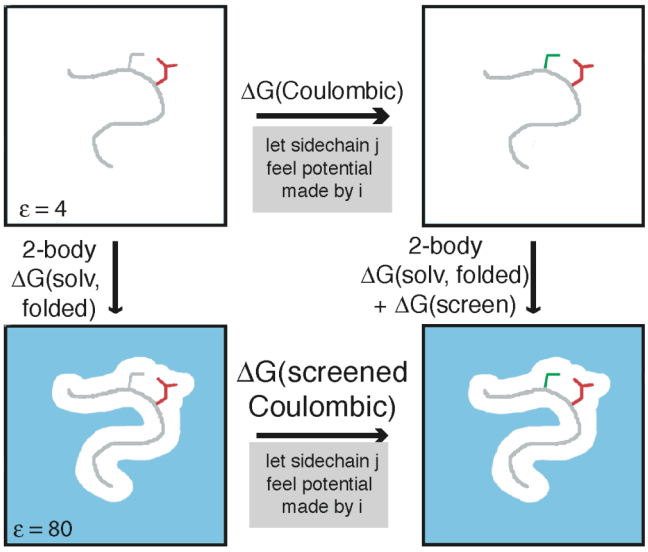

A) Exact sidechain desolvation & sidechain / backbone screened Coulombic energy
ε = 4
ΔG(electrostatic) = 0
add backbone and ALL sidechains
ΔG(Coulombic)
let backbone feel potential made by i
ΔG(solv, unfolded)
ΔG(solv, folded)
ΔG(solv, folded) + ΔG(screen)
ε = 80
ΔG(desolv)
add backbone and ALL sidechains
ΔG(screened Coulombic)
let backbone feel potential made by i
B) One- & two-body sidechain desolvation
ε = 4
ΔG(electrostatic) = 0
add backbone ONLY
ΔG(electrostatic) = 0
add sidechain j
ΔG(solv, unfolded)
1-body ΔG(solv,folded)
2-body ΔG(solv,folded)
ε = 80
ΔG(desolv)
add backbone ONLY
ΔG(2-body perturbation)
add sidechain j
C) One- & two-body sidechain / backbone screened Coulombic energy
ε = 4
ΔG(Coulombic)
let backbone feel potential made by i
ΔG(electrostatic) = 0
add sidechain j
1-body ΔG(solv, folded)
1-body ΔG(solv, folded) + ΔG(screen)
2-body ΔG(solv, folded) + ΔG(screen)
ε = 80
ΔG(screened Coulombic)
let backbone feel potential made by i
ΔG(2-body perturbation)
add sidechain j

Figure 3-3. Free energy cycles used to calculate (A) exact versus (B) two-body sidechain/sidechain screened Coulombic energies (as shown in Equations 6–7 and 15–16, respectively). In each method, the electrostatic potential generated by sidechain *i* is multiplied by the charges in sidechain *j* to determine the screening energy between sidechain *i* and sidechain *j*. The key distinctions between the exact and two-body methods are as follows. The exact calculation uses the protein backbone and all of the sidechains in the protein to define the dielectric boundary, while the two-body calculation uses the protein backbone and only two sidechains to define the dielectric boundary. The parameters used in each FDPB calculation are indicated as follows: sidechain *i*, shown in red, was assigned partial atomic charges from the PARSE charge set; the rest of the protein, when shown in gray, was assigned partial atomic charges of 0; sidechain *j*, when shown in green, was assigned partial atomic charges of 0 in the FDPB calculation, but its PARSE partial atomic charges were used to obtain screening energies; the areas drawn in white were assigned a dielectric constant of 4 (protein interior); and the blue areas were assigned a dielectric constant of 80 (water) and a salt concentration of 50 mM.

A) Exact sidechain i / sidechain j screened Coulombic energy



B) Two-body sidechain i / sidechain j screened Coulombic energy



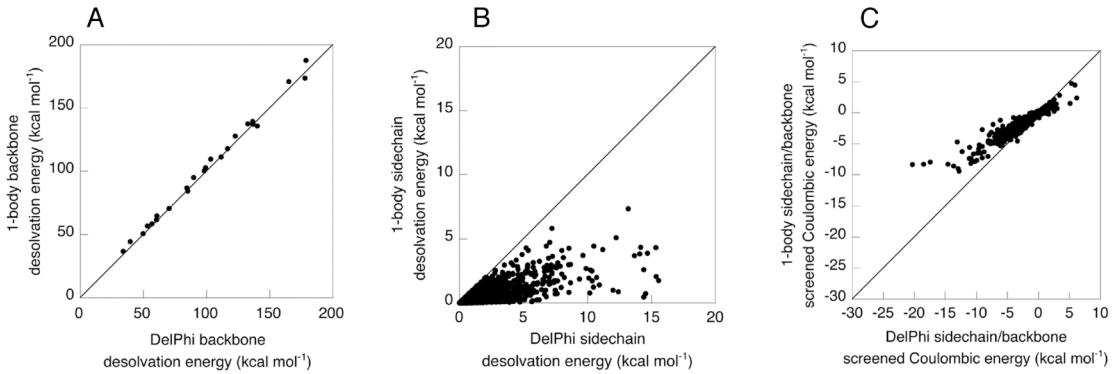


Figure 3-4. Accuracy of the one-body method determined by comparing (A) exact FDPB backbone desolvation energies versus one-body backbone desolvation energies, (B) exact FDPB sidechain desolvation energies versus one-body sidechain desolvation energies, and (C) exact FDPB sidechain/backbone screened Coulombic energies versus one-body sidechain/backbone screened Coulombic energies.

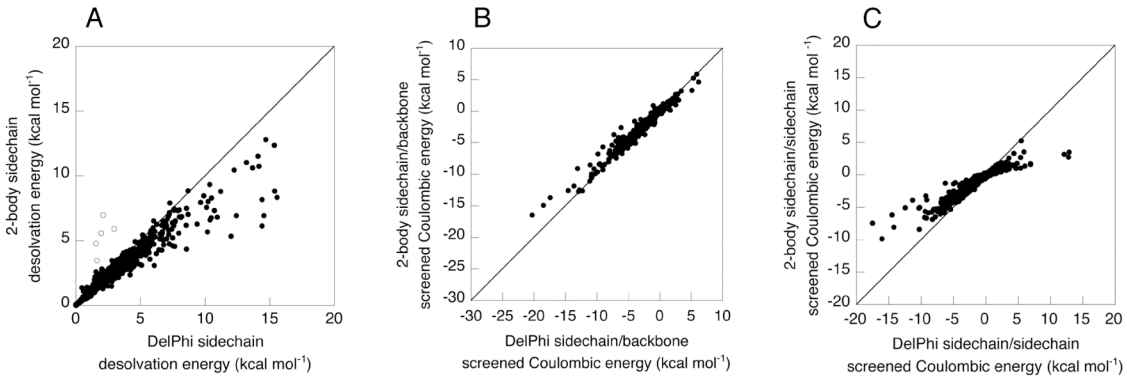


Figure 3-5. Accuracy of the two-body method determined by comparing (A) exact FDPB sidechain desolvation energies versus two-body sidechain desolvation energies with outlier points represented by open circles, (B) exact FDPB sidechain/backbone screened Coulombic energies versus two-body sidechain/backbone screened Coulombic energies, and (C) exact FDPB sidechain/sidechain screened Coulombic energies versus two-body sidechain/sidechain screened Coulombic energies.

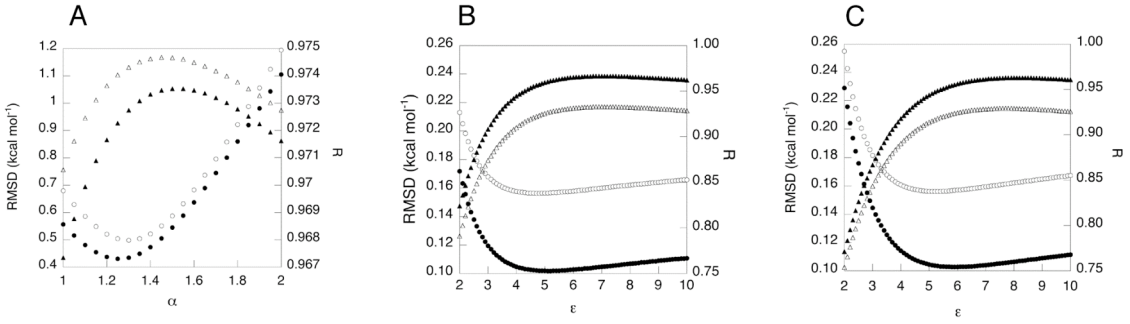


Figure 3-6. Sensitivity of error in two-body energies due to changes in (A) α, the scaling the parameter for two-body sidechain desolvation energies, (B) the distance dependent dielectric for pairs separated by greater than 6.0 Å, and (C) the distance dependent dielectric for pairs separated by greater than 4.0 Å. In all cases, filled symbols refer to protein structure set 1, open symbols refer to protein structure set 2, circles indicate RMSD, and triangles indicate the correlation coefficient R.

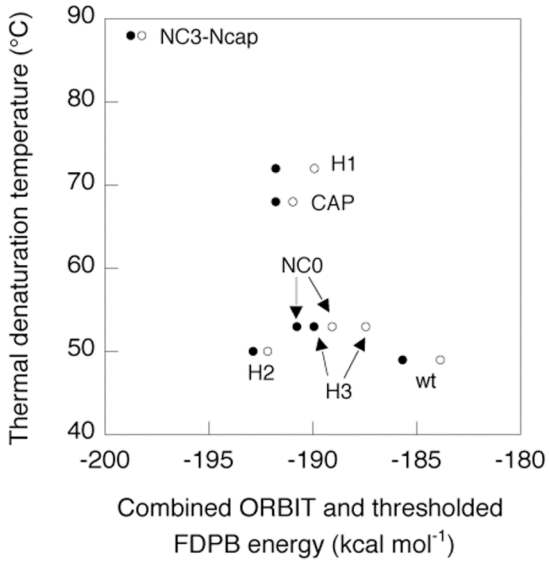


Figure 3-7. Energy predicted using the sum of the FDPB sidechain desolvation energy, FDPB sidechain/backbone screened Coulombic energy, FDPB sidechain/sidechain screened Coulombic energy and ORBIT van der Waals energy versus the experimentally determined stability of each homeodomain variant. The energies obtained using the two-body FDPB approximation are shown as filled circles, and the energies obtained using the exact FDPB model are shown as open circles.

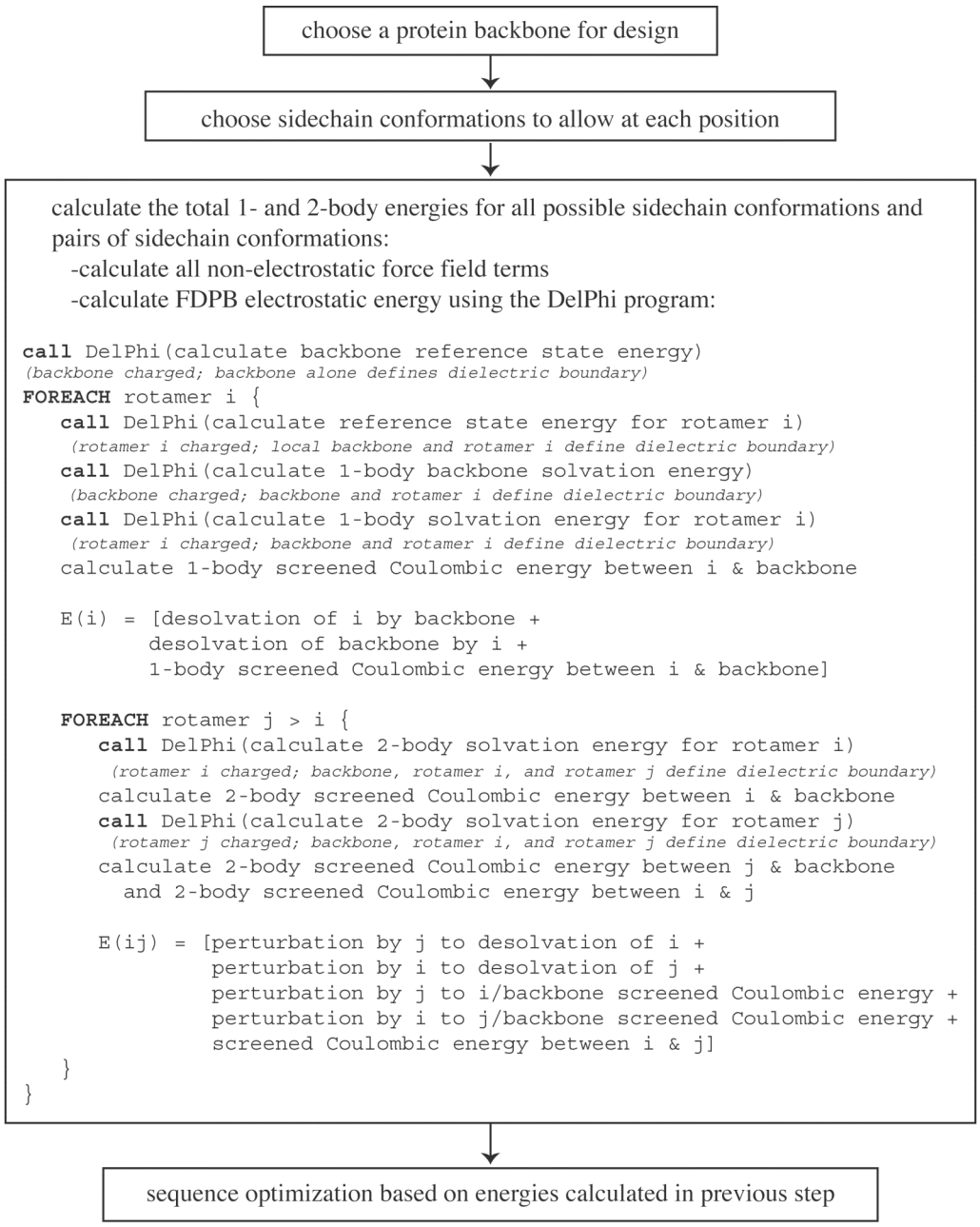


Figure 3-8. Protein design protocol, including one and two-body FDPB calculations. A simplified version of the protein design procedure shows the step at which electrostatic energies are calculated. Pseudocode for the electrostatics calculation shows the number of times in which the FDPB program DelPhi is called. For screened Coulombic energies, the potential maps from previous calculations are used to obtain screening energies. As described in the methods section, desolvation energies are computed as the difference between folded state and reference state solvation energies. Two-body perturbations for sidechain desolvation and sidechain/backbone screened Coulombic energies are computed as the difference between the respective two-body and one-body energies.

Chapter 4

An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design

The text of this chapter is adapted from a published manuscript that was co-authored with Naigong Zhang,[1] Shannon A. Marshall, Professor Ned S. Wingreen,[2] Professor Chen Zeng,[1] and Professor Stephen L. Mayo

C.L. Vizcarra, N. Zhang, S.A. Marshall, N.S. Wingreen, C. Zeng, and S.L. Mayo, *Journal of Computational Chemistry* **7**, 1153–1162 (2008).

[1]Department of Physics, George Washington University, Washington, DC

[2]Department of Molecular Biology, Princeton University, Princeton, NJ

Abstract

Our goal is to develop accurate electrostatic models that can be implemented in current computational protein design protocols. To this end, we improve upon a previously reported pairwise decomposable, finite difference Poisson-Boltzmann (FDPB) model for protein design.[1] The improvement involves placing generic sidechains at positions with unknown amino acid identity and explicitly capturing two-body perturbations to the dielectric environment. We compare the original and improved FDPB methods to standard FDPB calculations in which the dielectric environment is completely determined by protein atoms. The generic sidechain approach yields a two- to threefold increase in accuracy per residue or residue pair over the original pairwise FDPB implementation, with no additional computational cost. Distance-dependent dielectric and solvent-exclusion models were also compared to standard FDPB energies. The accuracy of the new pairwise FDPB method is shown to be superior to these models, even after re-parameterization of the solvent-exclusion model.

Introduction

Current computational protein design programs could be improved by the inclusion of an accurate model for electrostatics. Since proteins exist in highly polarizable solvents, the accuracy of the electrostatics model is dependent on the accuracy of the solvation model. To overcome the computational demands of explicitly modeling all of the water molecules in a macromolecular system, a continuum dielectric description of water is used in many biomolecular applications.[2,3] In continuum solvation models, the protein is treated as a low dielectric cavity within a high dielectric solvent. The boundary between the protein and solvent dielectric is defined by the protein's molecular surface. When carrying out amino acid sequence selection for protein design, the location of the dielectric boundary becomes ambiguous because the final amino acid identities and their conformations are not known until the very end of the calculation. In order to overcome this limitation and to satisfy the need for computationally efficient energy functions, alterations to the Generalized Born model,[4,5] a modified version of the Tanford-Kirkword model,[6,7] and various empirical models[8–11] have been reported for protein sequence design.

Within the limitations of a continuum solvent description, the Finite Difference Poisson Boltzmann (FDPB) model is often considered a standard for accuracy.[12,13] A general strategy for implementing an FDPB model that is pairwise decomposable by sidechain conformation (rotamer) has been reported.[1] This strategy involves evaluating explicit perturbations to the dielectric boundary. For example, the desolvation energy of a sidechain on being transferred from the unfolded state to the folded state is calculated by solving for the difference in solvation energy between the one-body state (i.e., the

folded backbone and one sidechain) and the unfolded state model for the sidechain. Two-body perturbations are calculated as the difference in solvation energy between a state with two sidechains (the "two-body state" in Fig. 4-1B) and the one-body state. The total pairwise sidechain desolvation is thus the desolvation of the sidechain by the backbone plus the sum of two-body perturbations. The energy terms in this method are fully pairwise decomposable by sidechain conformation and are therefore compatible with the energy matrices and optimization algorithms used in most computational design methods.

The accuracy of the pairwise decomposable FDPB model was assessed by comparing the energy calculated with the entire molecular surface defined by all of the protein sidechains (the "exact" surface in Fig. 4-1A) to the energy calculated using the sum of perturbations method. It was found that the desolvation of sidechains could be accurately approximated with an RMS error of 0.64 kcal mol^{-1} per sidechain.[1] The generic sidechains described by Zhang *et al.*[14] for calculation of pairwise solvent-accessible surface area present a straightforward and efficient strategy for improving the accuracy of pairwise approximate FDPB calculations. Figure 4-1 shows the difference between the original pairwise FDPB model and the generic sidechain approach. At all positions for which the identity or conformation of the amino acid is unknown, a generic sidechain composed of three spheres is placed, making the one-body state more closely resemble the true protein molecular surface and the two-body perturbations less dramatic.

A generic sidechain approach to approximating the volume occupied by a protein's sidechains has been used previously in many applications, including residue classification with respect to the molecular surface,[15,16] protein-protein docking,[17] and

solvation.[14] Pokala and Handel have reported a one-body generic sidechain formulation of the Generalized Born (GB) model.[4] For each residue in a design calculation, they approximate the low dielectric environment by spheres at all other positions. We take a similar approach using the FDPB model, but, importantly, we also calculate two-body perturbations that lead to a better approximation of the protein environment. In order to overcome the computational limitations of an $O(n^2)$ calculation, distance cutoffs are tested.

Due to the computational demands of solving the PB equation numerically, there is a great deal of interest in methods that approximate the PB model, such as the GB model, and also in fast empirical models.[2] The solvent-exclusion model of Lazaridis and Karplus (LK)[18] is computationally efficient and has been used by Baker and coworkers in the successful design of a novel fold.[19] Here we test the improved pairwise FDPB model against the LK model. Since the original parameterization of the LK model was based primarily on experimental solvation free energies, we derive new parameters based on FDPB energies to see how well the functional form of the LK model is able to reproduce this particular benchmark. While it is found that the generic sidechain method out-performs the LK model, the trade-off between computational efficiency and accuracy of the energy function is discussed.

Methods

FDPB calculations. A set of 24 proteins with hydrogens added was taken from the Richardson Top 500 database of high resolution X-ray crystal structures (http://kinemage.biochem.duke.edu/databases/top500.php). The PDB codes for the set

are: 1IGD, 1MSI, 1KP6, 1OPD, 1FNA, 1MOL, 2ACY, 1ERV, 1DHN, 1WHI, 3CHY, 1ELK, 2RN2, 1HKA, 3LZM, 1AMM, 1XNB, 153L, 1BK7, 2PTH, 1THV, 1BS9, 1AGJ, and 2BAA. The DelPhi program[20] was used to solve the linearized Poisson-Boltzmann equation using the following settings: 2 grids/Å, 0.05 M salt, a protein dielectric of 4 and a solvent dielectric of 80. In all calculations on a single structure, the protein's position relative to the grid was held constant. PARSE radii and charges were used.[21] The test set contains 2028 polar residues when using PARSE charge definitions. All prolines and disulphide bonds were treated as part of the backbone.

The three sphere generic sidechain method (herein referred to as G3) reported by Zhang et al.[14] was used in all calculations described below unless otherwise noted. Calculations denoted G0 refer to the method of Marshall et al.[1] A grid-based search was carried out to find a more optimal set of generic sidechain dimensions for FDPB calculations. Within the grid search, the parameters reported previously[14], sphere radius = 2.85 Å and distance between spheres = 0.61 Å, were found to be near-optimal and were used as given.

G3 parameter optimization. The radius of the generic sidechain spheres and the distance between those spheres (Figure 4-1) were varied to find optimal parameters for pairwise FDPB calculations. For the grid search of generic sidechain dimensions, a subset of 10 structures was used: 1IGD,1KP6, 1FNA, 2ACY, 1DHN, 3CHY, 2RN2, 3LZM, 1BK7, and 1THV. The results of those trials are given in Tables 4-5 and 4-6. The values in Table 4-5 correspond to the force field terms presented in Table 4-1. Since the error values in the various terms did not minimize at the same radius and distance, a more

general measure of error in the total reaction field energy was formulated. Using the linearized PB equation, the total solvation or reaction field energy for each structure in the folded state can be expressed as

$$\Delta G^{solv}_{total} = \Delta G^{solv}_{sidechain} + \Delta G^{solv}_{backbone} + \Delta G^{screen}_{sb} + \Delta G^{screen}_{ss}$$

where $\Delta G^{solv}_{backbone}$ is the backbone folded state solvation energy, $\Delta G^{solv}_{sidechain}$ is the total sidechain solvation energy for the structure, ΔG^{screen}_{sb} is the total screening energy between the backbone and sidechains, and ΔG^{screen}_{ss} is the total screening energy between all pairs of polar or charged sidechains. To minimize cancellation of errors, the error in each of these terms was assessed for each structure. The RMS error over the terms was evaluated for each structure and the sum of the RMS errors over the ten structures was used to assess the total error associated with each parameter set,

$$total\ error = \sum_{training\ set} \sqrt{\left(\frac{1}{4}\right)\left(\left(\Delta G^{solv,exact}_{sidechain} - \Delta G^{solv,G3}_{sidechain}\right)^2 + \left(\Delta G^{solv,exact}_{backbone} - \Delta G^{solv,G3}_{backbone}\right)^2 + \left(\Delta G^{screen,exact}_{sb} - \Delta G^{screen,G3}_{sb}\right)^2 + \left(\Delta G^{screen,exact}_{ss} - \Delta G^{screen,G3}_{ss}\right)^2\right)}.$$

The numerical results of this optimization are given in Table 4-6 and a plot of total error as a function of radius and distance is given in Figure 4-7.

Protein design energy terms. In order to be consistent with the ORBIT force field[22], backbone desolvation, sidechain desolvation, sidechain/backbone screened Coulombic, and sidechain/sidechain screened Coulombic energies were calculated. The one- and two-body calculations are analogous to those described in Marshall et al.[1] However, for all G3 calculations, three sphere generic sidechains were used at all positions for which no sidechain was present. The unfolded state reference for sidechain desolvation consisted of the sidechain *i* plus the local backbone atoms: CA(i-1), C(i-1), O(i-1), N(i), H(i), CA(i), C(i), O(i), N(i+1), H(i+1), and CA(i+1). Screened Coulombic interactions

were only calculated in the folded state. The most notable difference between the G0 and G3 methods is the calculation of the backbone desolvation energy (ΔG^{bb}_{desolv}). The unfolded state for the backbone is still described by a crystallographic backbone with no sidechains present. A "zero-body" folded state is then defined by the backbone with generic sidechains at all positions. Single residue (one-body) perturbations to the zero-body state are summed to get the total one-body backbone desolvation energy:

$$\Delta G^{bb}_{desolv} = \Delta G^{bb}_{zero-body} - \Delta G^{bb}_{unfolded} + \sum_{i}^{n}\left(\Delta G^{bb,i}_{one-body} - \Delta G^{bb}_{zero-body}\right). \qquad (1)$$

By Equation 1, the backbone desolvation energy is derived from (n + *2*) DelPhi calculations, where n is the number of residues in the protein. In order to calculate all of the one- and two-body energies for a structure with n residues, p of which are polar, a total of

$$(n+2) + 2p + p(n-p) + p(p-1) \qquad (2)$$

DelPhi calculations are needed, where (n + *2*) corresponds to backbone desolvation, *2p* corresponds to the unfolded state and one-body folded state models, p(*n* - *p*) corresponds to perturbations of polar residues by non-polar residues, and p(*p* - *1*) corresponds to perturbations and interactions between polar residues, noting that two-body perturbations are not symmetric.

DDD and LK calculations. New parameters for the solvent-exclusion model originally reported by Lazaridis and Karplus (LK)[18] were derived using the following 14 structure training set: 1MSI, 1OPD, 1MOL, 1ERV, 1WHI, 1ELK, 1HKA, 1AMM, 1XNB, 153L, 2PTH, 1BS9, 1AGJ, and 2BAA. While the CHARMM19 parameters described by Lazaridis and Karplus[18] are atom based, the new parameters are sidechain based. For

instance, lysine was assigned a single parameter for all heavy atoms with non-zero partial atomic charges in the PARSE charge set: C_ε and N_ζ. Since the desolvation energy of a sidechain in the LK model is independent of the ΔG_{ref} parameter, only values for ΔG_{free} were derived by fitting to the "exact" FDPB desolvation energy

$$\Delta G^i_{desolv} = \Delta G^i_{free} \sum_{t \in i} \sum_{\substack{u \notin i, \\ local\ bb}} f_t(r_{tu}) V_u \tag{3}$$

where the function f is the Gaussian free energy density of atom t and V_u is the volume of desolvating atom u. For each sidechain i in the training set, the sum of Gaussian solvent exclusion terms was calculated over each atom t in sidechain i and each atom u that is not in the sidechain i or its local backbone. For each set of amino acids, with the following amino acids considered together: Asn and Gln; Ser and Thr; and Asp and Glu, a linear least-squares fit was used to get ΔG_{free} from the exact FDPB sidechain desolvation energy ΔG_{dsolv}. The new LK parameters are listed in Table 4-4. For the distance-dependent dielectric (DDD) calculations, dielectrics were assigned as 5.1r for sidechain/backbone interactions and 7.1r for sidechain/sidechain interactions, according to Zollars et al.[11] These values were derived by fitting to FDPB screened Coulombic energies. Since there were five structures in common between their training set and the 24 structures used here, the error in screened Coulombic energy was assessed for the remaining 19 structures.

Results

The accuracy of the pairwise FDPB methods was measured by comparison to "exact" FDPB energies calculated with all protein atoms present. The results of this comparison are shown for both the G0 and G3 models in Table 4-1 and Figures 4-2 and 4-3. The G3 model performs better than the G0 model in all cases. As expected, the one-body sidechain desolvation improves dramatically over the G0 model in which only desolvation of the sidechain by the backbone is counted (Figures 4-2C and 4-2D). Similarly, the one-body G3 model is more effective than the one-body G0 model at capturing the descreening of strong sidechain/backbone interactions (Figures 4-2E and 4-2F). It is interesting to note that the one-body G3 model for sidechain desolvation is more accurate than the two-body G0 model with a 4 Å cutoff (Table 4-1). This indicates that the approximate surface provided by the generic sidechains is more effective at reproducing the exact energies than adding the one-body energy to the truncated sum of two-body sidechain perturbations in the G0 model, a relevant model to consider since a distance cutoff will almost certainly be used in design calculations.

As shown in Table 4-1 and Figure 4-3, the G3 two-body models for sidechain desolvation and sidechain/backbone screened Coulombic energy provide improvements over the one-body models and the G0 two-body models. An especially dramatic improvement in accuracy is seen for the two-body approximation for sidechain/sidechain interactions. Each data point in Figures 4-3E and 4-3F corresponds to a pair of residues. For the G0 model, there are no descreening contributions from other sidechains to sidechain/sidechain interactions, whereas the generic sidechains provide a vast improvement by approximating the reduced dielectric of other sidechains. The accuracy

of the model is only slightly reduced by using an approximate distance-dependent dielectric model for pairs separated by more than the specified cutoffs. The dielectric values were based on those reported by Marshall et al.[1] For both cutoff values tested, more than 90% of all polar sidechain pairs in the test set were not treated with an FDPB calculation. Such cutoffs would provide a considerable speed enhancement in the energy calculation stage of a design calculation.

Although the G3 model performs better than the G0 model for sidechain desolvation, there are noticeable outliers in Figure 4-3B. There was also one residue in the test set with a negative G3 two-body desolvation energy which is not shown in Figure 4-3B but is included in the calculated error in Table 4-1. Out of the 2028 residues in the test set, there are 19 for which sidechain desolvation is underestimated by more than 1.5 kcal mol^{-1} when using the G3 model. For this set of 19 residues, the amino acid types are exclusively Asp, Glu, Arg, and Lys, and they are from 12 different structures. Two of these outliers, shown as "✕" symbols in Figure 4-3B, plus the point with a negative desolvation were sensitive to moving the molecule slightly with respect to the grid. This sensitivity to grid placement has been discussed previously for the pairwise FDPB calculation[1] and for more standard applications.[23] An additional seven of the outliers had nearby residues that gave large, negative perturbations, leading to two-body energies with larger error than the one-body approximation. In five of the seven cases, these large negative perturbations are caused by glycines, the amino acid for which the G3 approximation is the most inaccurate. The remaining ten cases with large, negative error had exact desolvation energies greater than 8 kcal mol^{-1} and one-body error greater than the two-body error, indicating that these points are simply difficult to capture by a

pairwise summation scheme. For both two-body sidechain desolvation and sidechain/backbone screened Coulombic energy, the error decreases slightly when cutoffs are imposed. This may point to the fact that only local perturbations are necessary with the G3 model and that inclusion of longer range perturbations leads to errors in accounting for sidechain overlap, as described by Zhang et al.[14]

The generic sidechain parameters used here are the same as those used in Zhang et al.[14] for solvent accessible surface area calculations. Since an alternative set of parameters may be more optimal for the molecular surface definition used in FDPB calculations, we carried out a grid search of parameters to find a superior set of parameters and to assess how sensitive the G3 method is to generic sidechain dimensions. As shown in Figures 4-4 and 4-5, the error is sensitive to sphere size and spacing over the entire parameter space explored, but parameter sets near radius = 2.85 Å and distance = 0.61 Å have relatively low error. The optimality of these parameters for both surface area and FDPB calculations supports the assertion that generic sidechains of these dimensions accurately represent the average space occupied by amino acid sidechains in folded proteins.[24] Therefore, if the surface definition (e.g., solvent-accessible or molecular surface) is consistent between the pairwise and exact calculations, these parameters will be suitable. It is also notable that the training set with which these parameters were originally derived has an overlap of only one protein with the 10 structures used in the parameter search here, suggesting that these parameters are robust for general protein design targets.

It is of general interest to see how the pairwise approximate FDPB method described here performs in comparison to fast pairwise decomposable methods already

used for protein design.[2] We compared the performance of the solvent-exclusion model of Lazaridis and Karplus (LK)[18] and a distance-dependent dielectric (DDD) model with the G3 method. Both of the LK and DDD models are highly parameterized. Since multiple parameter sets have been reported for the LK model, we tried both the CHARMM19 LK parameters[18] and a new set of parameters tuned specifically to reproduce PB energies. We used the distance-dependent dielectric values that led to a stabilized designed protein in a recent experimental protein design study.[11] The results of this comparison are shown in Table 4-2 and Figures 4-5 and 4-6. The G3 model is more effective than the LK and DDD models at approximating exact FDPB calculations. While the performance of the LK model (Figures 4-5B and 4-5C) varies greatly with parameters, the LK model has a nonlinear relationship with exact PB desolvation energies regardless of parameter set.

Discussion

We have shown that it is possible to improve the agreement between a pairwise decomposable FDPB method and an exact FDPB method with no additional computational cost. The improvement stems from the more accurate approximation of the dielectric boundary provided by generic sidechains. The RMS errors for screened Coulombic interactions between polar sidechains and between polar sidechains and the protein backbone are decreased by nearly threefold and twofold, respectively. The error associated with the desolvation of polar sidechains is reduced by nearly twofold. For the two-body perturbation-based terms (i.e., sidechain desolvation and sidechain/backbone screened Coulombic energy), this more accurate description of the dielectric boundary

leads to less dramatic perturbations to that boundary, accounting for the inherent non-additivity of such perturbations more effectively than the previously reported pairwise FDPB method.[1]

Ideally, a protein design energy function is both accurate and computationally efficient. The FDPB methods are approximately four orders of magnitude slower than the standard ORBIT energy function. For example, a surface design of the small 51-residue helical protein engrailed homeodomain gave 6.4 million rotamer pairs for 29 design positions. The pre-calculation of all rotamer singles and pairs energies required on the order of ~ 0.2 CPU hours using the standard ORBIT energy function with a surface-area-based solvation term and ~ 1000 CPU hours using the G3 model. This large computational cost requires one to carefully assess the appropriateness of the G3 model for different design problems. Because of the large investment of time and resources involved in synthesizing and characterizing designed proteins, an expensive calculation may be worthwhile, especially if the calculation involves positions with important electrostatic contacts such as in the active site of an enzyme. Unfortunately, the cost of the FDPB models may preclude large design targets since the calculation also scales poorly with the size of the grid on which the PB equation is solved.

The results shown here indicate good agreement between standard many-body electrostatic energies and those derived from summing one- and two-body perturbations. Standard FDPB calculations serve as a reasonable benchmark for assessing the sequence energies that will be evaluated by a pairwise decomposable force field. For search algorithms such as Monte Carlo[25] or FASTER[26], where total sequence energy is evaluated and used as a criteria to accept or reject rotamer changes, this benchmark is

sufficient. However, the comparison with standard FDPB calculations leaves the possibility of a cancellation of error when summing over perturbations: some perturbations may be "too small" and some may be "too large." This distinction may become important when using algorithms based on Dead End Elimination.[27] In such algorithms the sequence energy is not evaluated, but instead two-body perturbation energies are used to eliminate rotamers. There is no clear way to gauge the accuracy of the individual perturbation energies when comparing to standard benchmarks. Indeed, the most stringent test of this improved electrostatics term will be in the context of a protein design force field and, ultimately, in experimental validation of designed protein sequences.

References

1. Marshall, S. A., Vizcarra, C. L. & Mayo, S. L. (2005). One– and two–body decomposable Poisson–Boltzmann methods for protein design calculations. *Protein Sci.* **14**, 1293–1304.

2. Koehl, P. (2006). Electrostatics calculations: latest methodological advances. *Curr. Opin. Struct. Biol.* **16**, 142–151.

3. Vizcarra, C. L. & Mayo, S. L. (2005). Electrostatics in computational protein design. *Curr. Opin. Chem. Biol.* **9**, 622–626.

4. Pokala, N. & Handel, T. M. (2004). Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **13**, 925–936.

5. Archontis, G. & Simonson, T. (2005). A residue–pairwise generalized Born scheme suitable for protein design calculations. *J. Phys. Chem. B* **109**, 22667–22673.

6. Havranek, J. J. & Harbury, P. B. (1999). Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. USA* **96**, 11145–11150.

7. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10**, 45–52.

8. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* **316**, 189–199.

9. Wisz, M. S. & Hellinga, H. W. (2003). An empirical model for electrostatic interactions in proteins incorporating multiple geometry–dependent dielectric constants. *Proteins* **51**, 360–377.

10. Cerutti, D. S., Jain, T. & McCammon, J. A. (2006). CIRSE: A solvation energy estimator compatible with flexible protein docking and design applications. *Protein Sci.* **15**, 1579–1596.

11. Zollars, E. S., Marshall, S. A. & Mayo, S. L. (2006). Simple electrostatic model improves designed protein sequences. *Protein Sci.* **15**, 2014–2018.

12. Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A. & Brooks, C. L. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **25**, 265–284.

13. Baker, N. A. (2005). Improving implicit solvent simulations: a Poisson–centric view. *Curr. Opin. Struct. Biol.* **15**, 137–143.

14. Zhang, N. G., Zeng, C. & Wingreen, N. S. (2004). Fast accurate evaluation of protein solvent exposure. *Proteins* **57**, 565–576.

15. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87.

16. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* **305**, 619–631.

17. Huang, P. S., Love, J. J. & Mayo, S. L. (2005). Adaptation of a fast Fourier transform–based docking algorithm for protein design. *J. Comput. Chem.* **26**, 1222–1232.

18. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133–152.

19. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic–level accuracy. *Science* **302**, 1364–1368.

20. Rocchia, W., Alexov, E. & Honig, B. (2001). Extending the applicability of the nonlinear Poisson–Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem. B* **105**, 6507–6514.

21. Sitkoff, D., Sharp, K. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978–1988.

22. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513.

23. Gilson, M. K., Sharp, K. & Honig, B. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.* **9**, 327–335.

24. Creighton, T. E. (1993). *Proteins: Structure and Molecular Properties*. W.H. Freeman and C., New York.

25. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation Of State Calculations By Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092.

26. Desmet, J., Spriet, J. & Lasters, I. (2002). Fast and Accurate Side–Chain Topology and Energy Refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**, 31–43.

27. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead–end elimination theorem and its use in protein side–chain positioning. *Nature* **356**, 539–542.

Table 4-1: Accuracy of the electrostatic models

	G0[a]		G3[b]	
	RMSD (kcal mol^{-1})	R	RMSD (kcal mol^{-1})	R
A. Backbone desolvation energy				
one-body	3.96	0.997	3.51	0.998
B. Sidechain desolvation energy				
one-body	1.93	0.718	0.79	0.915
two-body, all pairs	0.64	0.962	0.40	0.979
two-body, pairs < 6 Å	0.67	0.968	0.35	0.984
two-body, pairs < 4 Å	0.82	0.952	0.39	0.980
C. Sidechain/backbone screened Coulombic energy				
one-body	0.90	0.957	0.34	0.987
two-body, all pairs	0.36	0.987	0.18	0.996
two-body, pairs < 6 Å	0.41	0.984	0.17	0.996
two-body, pairs < 4 Å	0.51	0.979	0.23	0.994
D. Sidechain/sidechain screened Coulombic energy				
two-body, all pairs	0.13	0.948	0.05	0.987
two-body, pairs < 6 Å[c]	0.13	0.939	0.06	0.979
two-body, pairs < 4 Å[d]	0.13	0.933	0.07	0.972

[a] Marshall et al 2005[1]

[b] Sphere radius = 2.85 Å, distance from C_α and distance between spheres = 0.61 Å

[c] For pairs separated by more than 6 Å, a distance dependent dielectric constant of 4.93r was used.

[d] For pairs separated by more than 4 Å, a distance dependent dielectric constant of 5.56r was used.

Table 4-2: Parameter sensitivity of the G3 model

	grid spacing	*2 grids Å^{-1}*	*2 grids Å^{-1}*	*2 grids Å^{-1}*	*4 grids Å^{-1}* [c]
	ionic strength	*50 mM*	*150 mM*	*50 mM*	*50 mM*
	# translations	*1*	*1*	*3*	*1*
one-body backbone desolvation		3.51	3.57	3.30	3.13
one-body sidechain desolvation		0.79	0.81	0.79	0.80
two-body sidechain desolvation[b]		0.40	0.41	0.38	0.41
one-body sidechain/backbone screened Coulombic energy		0.34	0.34	0.34	0.34
two-body sidechain/backbone screened Coulombic energy[b]		0.18	0.18	0.19	0.18
two-body sidechain/sidechain screened Coulombic energy[b]		0.05	0.06	0.05	0.06

[a]Error is reported as RMSD in kcal mol^{-1}.
[b]All two-body calculations were carried out without distance cutoffs.
[c]Only the exact calculations were carried out with a grid spacing of 4 grids Å^{-1}, while the one- and two-body calculations were carried with a grid spacing of 2 grids Å^{-1}.

Table 4-3: Comparison of FDPB, LK, and DDD models

	RMSD (kcal mol^{-1})	R
Sidechain desolvation energy[a,b]		
two-body G0	0.53	0.969
two-body G3	0.36	0.979
LK (CHARMM19)[c]	1.90	0.897
LK (tuned)	0.73	0.914
Sidechain/backbone screened Coulombic energy[a,d]		
two-body G0	0.37	0.986
two-body G3	0.19	0.996
DDD, $\varepsilon = 5.1r$[e]	0.83	0.921
Sidechain/sidechain screened Coulombic energy[a,d]		
two-body G0	0.13	0.943
two-body G3	0.05	0.986
DDD, $\varepsilon = 7.1r$[e]	0.14	0.915

[a]All two-body calculations were carried out without distance cutoffs.
[b]Ten structures in test set: 1IGD,1KP6, 1FNA, 2ACY, 1DHN, 3CHY, 2RN2, 3LZM, 1BK7, 1THV
[c]Lazaridis and Karplus, 1999[18]
[d]19 structures in test set: 1MSI, 1KP6, 1OPD, 1FNA, 1MOL, 2ACY, 1ERV, 1DHN, 3CHY, 1ELK, 1HKA, 1XNB, 153L, 1BK7, 2PTH, 1THV, 1BS9, 1AGJ, 2BAA
[e]Zollars, et al. 2006[11]

Table 4-4: LK parameters derived from FDPB energies

amino acid types	ΔG_{free}	R_{fit}	polar atoms[a]
Arg	-3.014	0.919	C_δ, N_ε, C_ζ, **$N_{\eta 1}$, $N_{\eta 2}$**
Asn / Gln	-1.937	0.960	C_γ, $O_{\delta 1}$, $N_{\delta 2}$ / C_δ, $O_{\varepsilon 1}$, $N_{\varepsilon 2}$
Asp / Glu	-5.344	0.924	C_γ, **$O_{\delta 1}$, $O_{\delta 2}$** / C_δ, **$O_{\varepsilon 1}$, $O_{\varepsilon 2}$**
Cys	-1.755	0.687	S_γ
His (neutral)	-0.989	0.875	C_β, C_γ, $N_{\delta 1}$, $C_{\delta 2}$, $C_{\varepsilon 1}$, $N_{\varepsilon 2}$
His (protonated)	-2.333	0.898	C_β, C_γ, **$N_{\delta 1}$**, $C_{\delta 2}$, $C_{\varepsilon 1}$, **$N_{\varepsilon 2}$**
Lys	-4.904	0.877	C_ε, **N_ζ**
Met	-0.673	0.947	C_γ, S_δ, C_ε
Phe	-0.308	0.946	C_β, C_γ, $C_{\delta 1}$, $C_{\delta 2}$, $C_{\varepsilon 1}$, $C_{\varepsilon 2}$, C_ζ
Ser / Thr	-4.117	0.924	O_γ / $O_{\gamma 1}$
Trp	-0.602	0.897	C_β, C_γ, $C_{\delta 1}$, $C_{\delta 2}$, $N_{\varepsilon 1}$, $C_{\varepsilon 2}$, $C_{\varepsilon 3}$, $C_{\zeta 2}$, $C_{\zeta 3}$, $C_{\eta 2}$
Tyr	-0.622	0.872	C_β, C_γ, $C_{\delta 1}$, $C_{\delta 2}$, $C_{\varepsilon 1}$, $C_{\varepsilon 2}$, C_ζ, O_η

[a]Atom types with correlation length $\lambda = 6.0$ Å are shown in bold. All other atoms were assigned $\lambda = 3.5$ Å.

Table 4-5: Accuracy of generic method for varied parameters[a]

r (Å)	d (Å)	1-body back-bone desolvation	1-body side-chain desolvation	2-body side-chain desolvation	1-body sidechain/ backbone screened Coulombic energy	2-body sidechain/ backbone screened Coulombic energy	2-body sidechain/ sidechain screened Coulombic energy
1.80	0.30	4.12	1.58	0.50	0.68	0.25	0.10
1.80	0.40	4.24	1.52	0.47	0.64	0.23	0.09
1.80	0.60	4.23	1.39	0.42	0.56	0.21	0.09
1.80	0.80	4.33	1.26	0.39	0.50	0.21	0.08
1.80	1.00	3.25	1.14	0.36	0.47	0.22	0.07
1.80	1.20	1.90	1.04	0.38	0.44	0.20	0.07
1.90	0.40	3.74	1.48	0.45	0.60	0.22	0.09
1.90	0.60	3.21	1.33	0.43	0.52	0.28	0.08
1.90	0.80	2.81	1.20	0.37	0.47	0.19	0.08
1.90	1.00	1.41	1.07	0.41	0.43	0.34	0.07
2.00	0.30	3.27	1.50	0.46	0.61	0.23	0.09
2.00	0.40	2.74	1.43	0.45	0.56	0.22	0.09
2.00	0.60	1.96	1.27	0.39	0.48	0.24	0.08
2.00	0.80	1.57	1.13	0.35	0.43	0.19	0.07
2.00	1.00	1.33	0.99	0.39	0.39	0.24	0.06
2.10	0.40	1.96	1.37	0.42	0.52	0.21	0.09
2.10	0.60	1.78	1.20	0.36	0.44	0.19	0.08
2.10	0.80	1.60	1.05	0.33	0.39	0.19	0.07
2.10	1.00	1.90	0.93	0.42	0.35	0.20	0.06
2.20	0.30	2.05	1.39	0.42	0.53	0.21	0.09
2.20	0.40	2.28	1.31	0.43	0.48	0.32	0.08
2.20	0.60	1.81	1.13	0.35	0.40	0.19	0.07
2.20	0.80	1.96	0.98	0.32	0.35	0.19	0.06
2.20	1.00	2.73	0.86	0.33	0.32	0.19	0.05
2.20	1.20	2.54	0.81	0.33	0.30	0.18	0.05
2.30	0.50	2.99	1.14	0.36	0.40	0.20	0.07
2.30	0.70	3.44	0.98	0.35	0.34	0.33	0.06
2.30	0.90	3.48	0.84	0.32	0.31	0.20	0.05
2.30	1.10	3.96	0.78	0.33	0.29	0.17	0.05
2.40	0.30	4.43	1.26	0.38	0.44	0.19	0.08
2.40	0.50	4.01	1.07	0.54	0.36	0.25	0.07
2.40	0.70	4.35	0.91	0.32	0.32	0.19	0.06
2.40	0.90	4.56	0.79	0.31	0.30	0.18	0.05
2.40	1.10	4.08	0.76	0.35	0.30	0.20	0.05
2.50	0.50	5.78	1.00	0.34	0.33	0.18	0.06
2.50	0.60	6.66	0.91	0.31	0.31	0.17	0.06
2.50	0.70	6.58	0.84	0.32	0.30	0.21	0.05
2.50	0.80	5.90	0.79	0.32	0.30	0.19	0.05
2.50	0.90	5.98	0.75	0.33	0.30	0.20	0.05
2.50	1.00	5.89	0.75	0.36	0.31	0.20	0.04
2.50	1.10	5.31	0.76	0.35	0.32	0.19	0.04
2.60	0.30	7.29	1.11	0.37	0.35	0.20	0.07
2.60	0.40	7.00	1.01	0.44	0.32	0.22	0.06
2.60	0.60	6.14	0.85	0.39	0.30	0.21	0.05
2.60	0.80	5.18	0.75	0.42	0.31	0.23	0.05

2.60	1.00	4.03	0.75	0.42	0.34	0.24	0.04
2.60	1.20	3.52	0.84	0.43	0.37	0.23	0.05
2.70	0.60	5.34	0.80	0.32	0.31	0.17	0.05
2.70	0.80	3.98	0.73	0.34	0.34	0.16	0.04
2.70	1.00	3.11	0.78	0.39	0.37	0.19	0.04
2.75	0.40	5.54	0.91	0.35	0.30	0.18	0.06
2.75	0.55	4.78	0.81	0.32	0.31	0.18	0.05
2.75	0.57	4.44	0.79	0.32	0.31	0.17	0.05
2.75	0.59	4.04	0.78	0.33	0.31	0.18	0.05
2.75	0.61	3.77	0.77	0.33	0.32	0.18	0.05
2.75	0.80	2.86	0.73	0.36	0.36	0.16	0.04
2.80	0.30	5.28	0.97	0.36	0.30	0.19	0.06
2.80	0.40	4.66	0.88	0.38	0.30	0.27	0.06
2.80	0.60	3.00	0.76	0.33	0.33	0.20	0.05
2.80	0.80	2.20	0.74	0.36	0.38	0.19	0.04
2.80	1.00	1.74	0.83	0.39	0.43	0.21	0.04
2.85	0.40	3.75	0.85	0.35	0.30	0.19	0.05
2.85	0.57	2.44	0.76	0.38	0.34	0.18	0.05
2.85	0.59	2.24	0.75	0.34	0.34	0.18	0.05
2.85	0.61	1.94	0.74	0.36	0.35	0.18	0.05
2.85	0.63	1.93	0.74	0.36	0.35	0.18	0.04
2.85	0.65	1.95	0.74	0.36	0.36	0.18	0.04
2.85	0.80	1.28	0.75	0.39	0.40	0.18	0.04
2.90	0.60	2.02	0.74	0.36	0.36	0.19	0.04
2.90	0.80	1.68	0.76	0.38	0.42	0.18	0.04
2.90	1.00	2.55	0.91	0.43	0.49	0.20	0.05
3.00	0.30	3.42	0.86	0.35	0.31	0.18	0.05
3.00	0.40	4.04	0.79	0.33	0.34	0.17	0.05
3.00	0.60	4.86	0.74	0.37	0.41	0.19	0.04
3.00	0.80	4.30	0.83	0.40	0.49	0.20	0.04
3.00	1.00	5.22	1.00	0.41	0.55	0.20	0.05
3.00	1.40	8.00	1.45	0.51	0.65	0.24	0.08
3.20	0.30	13.56	0.79	0.37	0.38	0.23	0.05
3.20	0.60	13.51	0.82	0.37	0.52	0.20	0.05
3.20	1.00	14.51	1.22	0.41	0.67	0.21	0.07

[a] For each term, RMSD is given in kcal mol^{-1}.

Table 4-6: Total error optimization[a]

R (Å)	D (Å)	1igd	1kp6	1fna	2acy	1dhn	3chy	2rn2	3lzm	1bk7	1thv	total
0.00	0.00	2.33	5.45	2.36	4.19	7.30	5.99	11.73	19.12	12.59	10.00	81.04
1.80	0.30	1.97	4.67	2.65	4.71	6.72	6.56	9.92	17.06	10.34	10.17	74.76
1.80	0.40	2.00	4.38	2.65	4.22	5.36	6.12	8.97	15.94	8.87	9.42	67.93
1.80	0.60	1.65	3.90	2.09	3.36	4.30	5.22	8.10	13.72	8.99	7.57	58.88
1.80	0.80	1.60	3.69	1.41	2.74	3.19	4.54	7.28	11.75	8.00	6.90	51.09
1.80	1.00	1.50	3.00	1.42	2.54	2.29	4.07	6.64	10.12	5.86	5.57	43.03
1.80	1.20	1.25	2.64	1.21	2.56	2.21	4.13	5.24	9.46	7.78	4.34	40.79
1.90	0.40	1.77	4.15	2.17	3.91	4.95	5.67	8.67	15.22	8.23	8.56	63.28
1.90	0.60	1.53	3.99	1.62	2.66	4.03	4.70	7.18	12.31	5.51	7.06	50.58
1.90	0.80	1.46	3.45	1.31	2.32	2.86	3.70	6.63	11.24	5.88	5.62	44.46
1.90	1.00	1.32	2.59	1.17	2.40	2.27	3.41	6.13	9.23	3.35	4.62	36.48
2.00	0.30	1.87	4.40	2.25	4.33	5.31	5.85	8.44	15.41	8.65	8.47	64.98
2.00	0.40	1.60	4.28	1.69	3.49	4.86	5.13	7.77	13.94	7.95	9.62	60.33
2.00	0.60	1.42	3.74	1.28	2.32	3.58	4.27	6.45	11.63	6.76	4.97	46.41
2.00	0.80	1.33	3.30	1.12	1.87	2.44	3.28	5.37	9.51	5.67	5.10	38.96
2.00	1.00	1.15	2.64	1.12	2.40	1.87	3.14	4.70	8.64	5.92	3.48	35.07
2.10	0.40	1.36	4.07	1.51	2.99	4.74	4.72	6.83	13.49	7.72	7.16	54.58
2.10	0.60	1.13	3.57	1.29	1.81	3.29	4.10	5.90	10.22	6.65	5.23	43.19
2.10	0.80	1.00	3.06	1.21	1.78	2.00	2.74	5.04	9.23	5.26	4.83	36.15
2.10	1.00	1.55	2.52	1.28	2.42	1.73	3.55	3.78	8.29	6.00	6.53	37.63
2.20	0.30	1.35	4.04	1.50	3.01	5.02	4.81	7.02	13.60	8.14	7.13	55.64
2.20	0.40	1.04	3.83	1.43	2.33	4.26	4.27	6.21	11.73	7.67	6.71	49.47
2.20	0.60	0.74	3.68	1.14	1.42	3.12	3.86	5.42	8.93	6.22	4.64	39.17
2.20	0.80	0.69	2.76	1.10	1.42	1.89	2.78	4.42	8.14	5.39	4.31	32.88
2.20	1.00	0.90	2.43	0.77	2.37	1.79	3.04	3.17	7.80	6.66	4.25	33.19
2.20	1.20	0.99	1.92	1.18	2.77	2.10	3.93	3.19	7.42	7.12	5.16	35.77
2.30	0.50	0.93	3.74	1.10	1.06	3.74	4.13	5.53	8.98	6.54	6.45	42.20
2.30	0.70	1.13	2.89	0.96	1.49	2.67	2.99	4.68	7.71	11.04	4.67	40.21
2.30	0.90	1.14	2.76	0.93	2.11	2.54	2.77	3.51	6.41	6.66	4.41	33.25
2.30	1.10	1.10	1.88	1.02	2.44	2.94	3.75	2.91	6.61	7.45	5.64	35.72
2.40	0.30	1.07	4.00	1.26	2.55	4.77	4.48	6.02	11.48	7.82	5.70	49.15
2.40	0.50	1.22	3.34	1.23	1.32	3.64	4.13	5.58	8.74	6.49	4.32	40.01
2.40	0.70	1.38	2.69	0.91	1.94	2.64	2.49	5.53	6.53	6.03	4.35	34.49
2.40	0.90	1.40	2.21	1.05	2.38	2.33	3.21	4.08	5.85	7.24	4.73	34.48
2.40	1.10	1.59	1.61	0.96	2.46	2.24	3.55	4.00	7.20	7.21	5.67	36.49
2.50	0.50	1.50	3.11	1.22	1.86	3.52	3.47	5.86	7.92	7.12	4.75	40.32
2.50	0.60	2.48	2.76	1.29	1.80	2.28	2.94	6.05	7.45	7.14	6.56	40.73
2.50	0.70	1.85	2.54	0.90	2.33	2.19	3.18	5.81	6.33	8.56	6.02	39.72
2.50	0.80	1.72	2.20	0.85	2.59	2.32	3.08	5.04	6.34	7.70	4.42	36.25
2.50	0.90	1.62	1.97	0.95	2.80	2.54	2.94	4.63	6.87	8.05	5.79	38.15
2.50	1.00	1.93	1.65	0.71	3.06	2.24	3.89	4.47	8.66	7.56	5.33	39.49
2.50	1.10	1.89	1.41	0.60	3.03	2.39	4.30	3.54	8.96	7.35	3.27	36.75
2.60	0.30	1.70	3.50	1.63	2.73	4.09	4.62	5.84	9.63	8.64	7.04	49.42
2.60	0.40	1.88	3.24	1.76	2.08	3.40	4.00	5.93	10.14	8.10	6.92	47.44
2.60	0.60	1.89	2.73	1.18	2.36	2.00	3.56	5.15	8.73	7.95	3.88	39.41
2.60	0.80	2.01	2.09	1.12	2.55	2.08	4.17	3.83	9.50	8.26	3.96	39.55
2.60	1.00	2.19	1.48	0.54	2.94	1.87	4.29	3.65	10.54	6.73	3.39	37.60
2.60	1.20	2.33	1.15	1.74	2.71	1.89	4.82	1.61	10.75	5.67	3.52	36.18
2.70	0.60	1.55	2.26	1.06	2.37	2.03	3.57	3.61	5.21	7.79	3.09	32.54

2.70	0.80	1.76	1.63	1.10	2.71	1.60	4.10	2.42	4.90	7.34	2.60	30.16
2.70	1.00	1.84	1.18	1.08	2.83	1.47	4.79	1.69	7.04	6.42	3.01	31.35
2.75	0.40	1.08	2.73	1.62	2.21	2.43	3.61	4.61	6.35	8.56	3.48	36.68
2.75	0.55	0.89	2.25	1.07	2.15	2.09	3.23	3.78	5.16	6.97	2.81	30.39
2.75	0.57	0.94	2.20	0.94	2.09	1.95	3.24	3.24	4.58	7.17	2.78	29.12
2.75	0.59	0.84	2.13	0.71	2.14	1.91	3.28	3.38	3.96	6.57	2.77	27.69
2.75	0.61	0.86	2.07	0.58	2.07	1.79	3.26	3.25	3.81	6.58	2.63	26.90
2.75	0.80	1.15	1.50	1.22	2.47	1.27	4.44	2.53	5.63	6.08	2.55	28.84
2.80	0.30	0.85	3.03	1.73	1.96	2.72	3.70	4.57	6.63	8.11	3.38	36.68
2.80	0.40	0.76	2.48	1.47	2.03	2.10	3.22	3.68	5.54	7.07	5.13	33.47
2.80	0.60	0.95	1.87	0.61	1.96	1.66	3.77	2.95	3.14	5.90	2.59	25.40
2.80	0.80	1.08	1.40	1.11	2.32	1.61	4.14	2.40	5.33	5.47	2.19	27.04
2.80	1.00	1.28	0.94	0.99	2.28	1.42	4.47	1.05	7.28	5.29	4.28	29.26
2.85	0.40	0.59	2.37	0.88	1.84	1.70	3.14	3.63	5.12	6.55	2.95	28.77
2.85	0.57	0.76	1.80	0.64	1.88	1.15	3.11	3.73	2.95	5.09	2.95	24.05
2.85	0.59	0.87	1.73	0.58	1.85	1.19	3.32	2.79	2.81	5.16	2.92	23.22
2.85	0.61	0.89	1.69	0.70	1.81	1.16	3.24	2.58	3.68	4.72	3.41	23.87
2.85	0.63	0.74	1.64	0.79	1.84	1.22	3.45	2.33	3.57	5.08	3.24	23.90
2.85	0.65	0.75	1.58	0.80	1.81	1.18	3.66	2.46	3.57	4.99	3.55	24.34
2.85	0.80	1.00	1.25	0.94	2.18	1.30	4.40	1.82	5.46	5.02	3.33	26.69
2.90	0.60	0.64	1.64	0.77	1.64	0.66	3.31	2.85	3.70	5.22	3.88	24.30
2.90	0.80	1.21	1.18	1.11	2.18	1.22	4.00	2.13	5.63	4.60	3.66	26.93
2.90	1.00	1.59	1.05	1.00	2.02	2.08	4.40	1.91	6.75	5.70	4.91	31.42
3.00	0.30	0.60	2.84	1.26	1.74	1.73	2.30	4.93	5.39	3.53	4.01	28.32
3.00	0.40	0.67	2.46	1.52	1.79	2.31	3.83	4.73	4.85	3.99	4.02	30.16
3.00	0.60	1.32	1.92	1.56	2.01	1.12	3.92	4.59	4.89	4.41	6.55	32.30
3.00	0.80	1.93	1.59	1.28	2.46	2.14	4.61	4.03	6.65	5.37	5.31	35.36
3.00	1.00	2.29	1.71	1.86	2.89	3.14	5.35	4.29	6.90	5.44	7.31	41.17
3.00	1.40	3.93	3.53	3.05	3.87	5.11	5.68	7.34	7.78	5.95	14.49	60.73
3.20	0.30	3.48	5.54	5.75	6.47	7.22	6.82	11.39	8.54	5.88	9.35	70.42
3.20	0.60	3.66	5.11	4.85	5.99	6.08	6.37	10.25	8.31	7.51	11.97	70.11
3.20	1.00	4.04	5.06	5.37	5.36	6.27	9.03	9.62	9.42	9.34	17.26	80.76

[a] Error is given in kcal mol^{-1}.

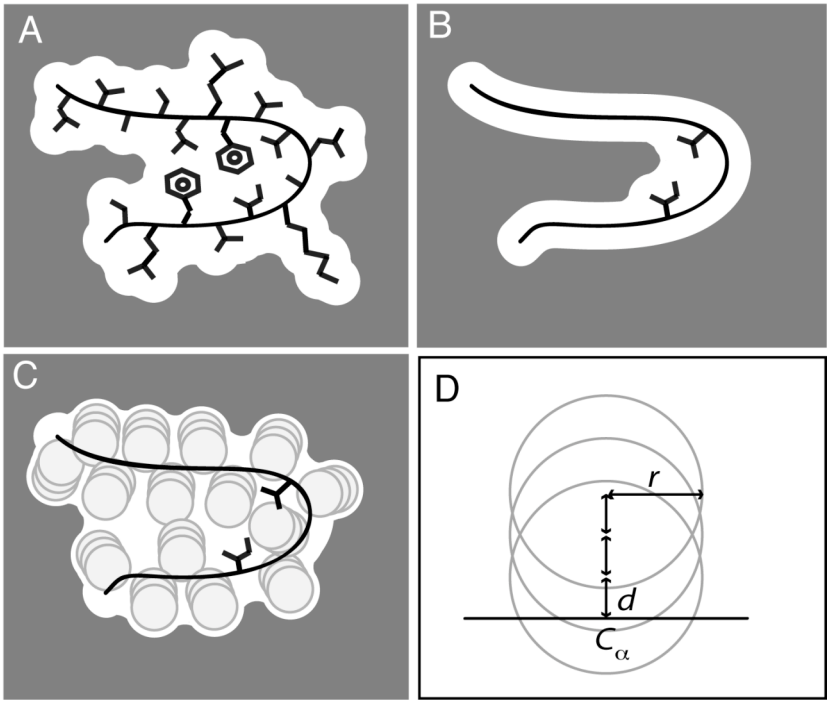


Figure 4-1. Illustration of exact, no generic sidechain (G0), and generic sidechain (G3) calculations. The dark gray area denotes solvent dielectric ($\varepsilon = 80$) and the white area denotes protein dielectric ($\varepsilon = 4$). (A) The exact molecular surface is defined by the backbone and all sidechains of the protein. (B) The two-body state for the G0 model is defined by two sidechains and the protein backbone. (C) The two-body state for the G3 model is defined by the two sidechains, the protein backbone, and three-sphere generic sidechains at all other positions. The one-body state is analogous to (B) or (C) but with only one sidechain represented explicitly. (D) The definition of radius and distance for the three-sphere generic sidechain is shown with sphere radius, *r*, and distance between spheres, *d*.

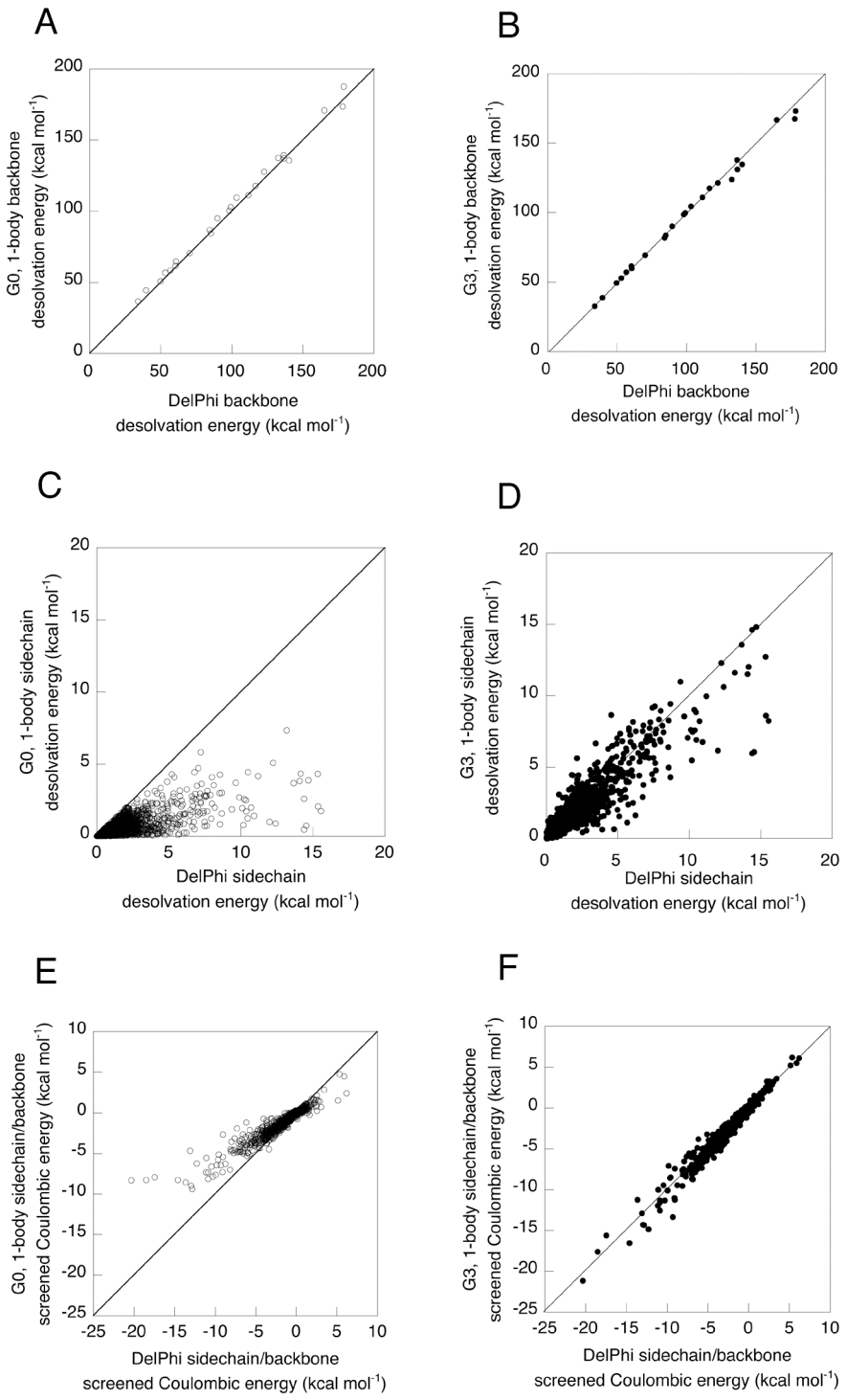


Figure 4-2. Accuracy of one-body G0 and G3 FDPB methods. One-body backbone desolvation calculated using the (A) G0 and (B) G3 methods. One-body sidechain desolvation calculated using the (C) G0 and (D) G3 methods. One-body screened Coulombic interaction energy between sidechains and backbone calculated using the (E) G0 and (F) G3 methods.

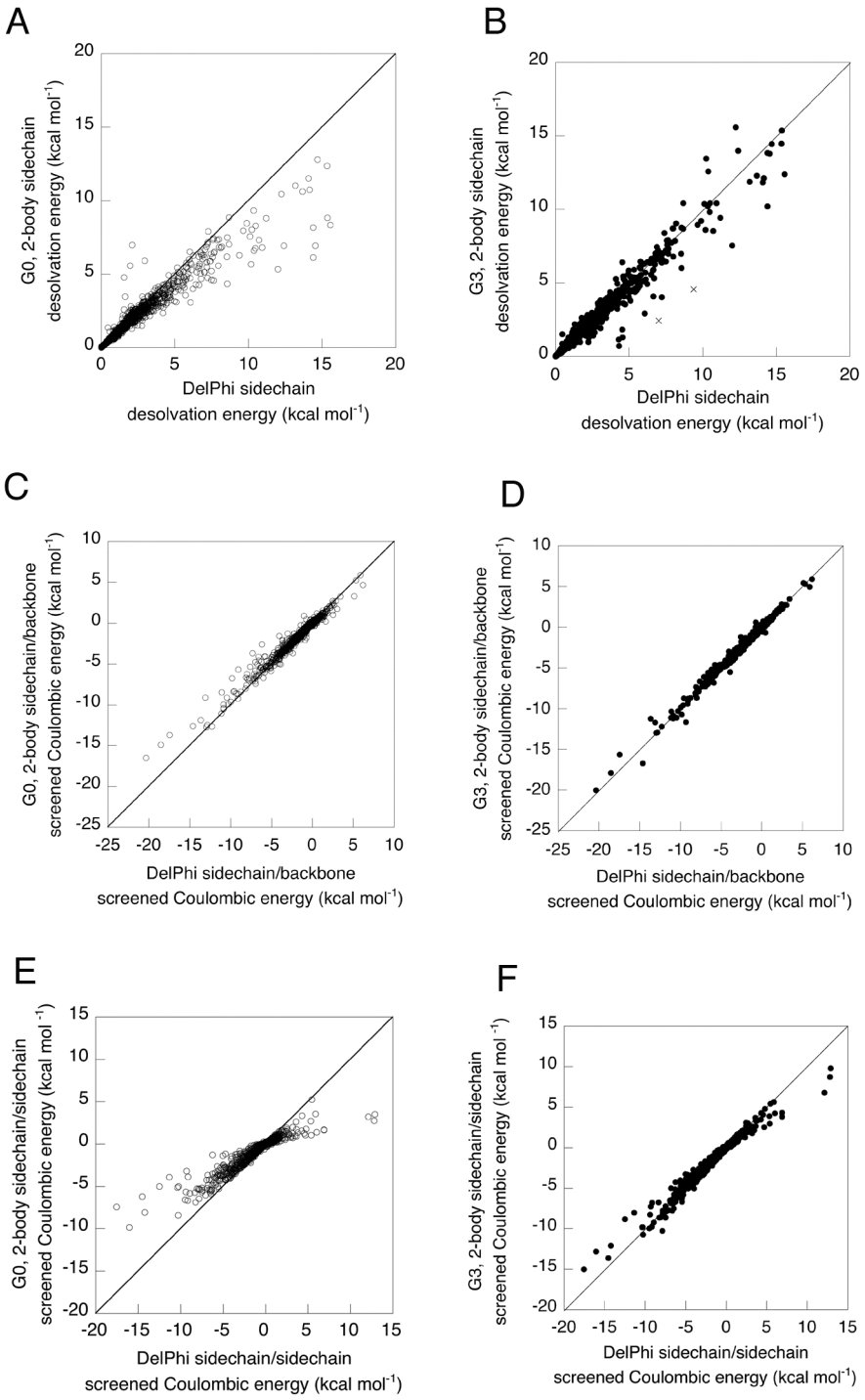


Figure 4-3. (see caption on next page)

Figure 4-3. Accuracy of two-body G0 and G3 FDPB methods. Two-body sidechain desolvation calculated using the (A) G0 and (B) G3 methods. Points marked with "×" in (B) correspond to sidechains for which the desolvation energy is sensitive to the placement of the protein with respect to the grid. Two-body screened Coulombic interaction energy between sidechains and backbone calculated using the (C) G0 and (D) G3 methods. Two-body screened Coulombic interaction energy between pairs of sidechains calculated using the (E) G0 and (F) G3 methods.

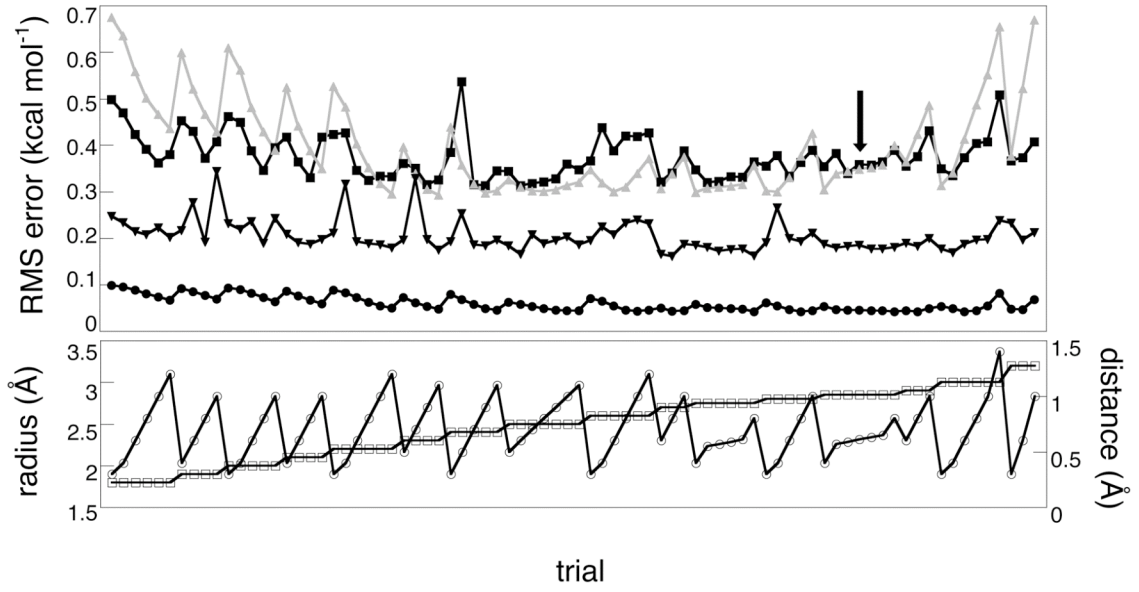


Figure 4-4. Sensitivity of the G3 FDPB method to generic sidechain parameters. Each line shows the error in a different force field component: two-body sidechain desolvation (■), one-body sidechain/backbone screened Coulombic energy (▲), two-body sidechain/backbone screened Coulombic energy (▼) , and two-body sidechain/sidechain screened Coulombic energy (●). The lower panel shows the radius (open squares) and distance (open circles) that were sampled in each trial. The parameter set radius = 2.85 Å and distance = 0.61 Å is indicated by an arrow.

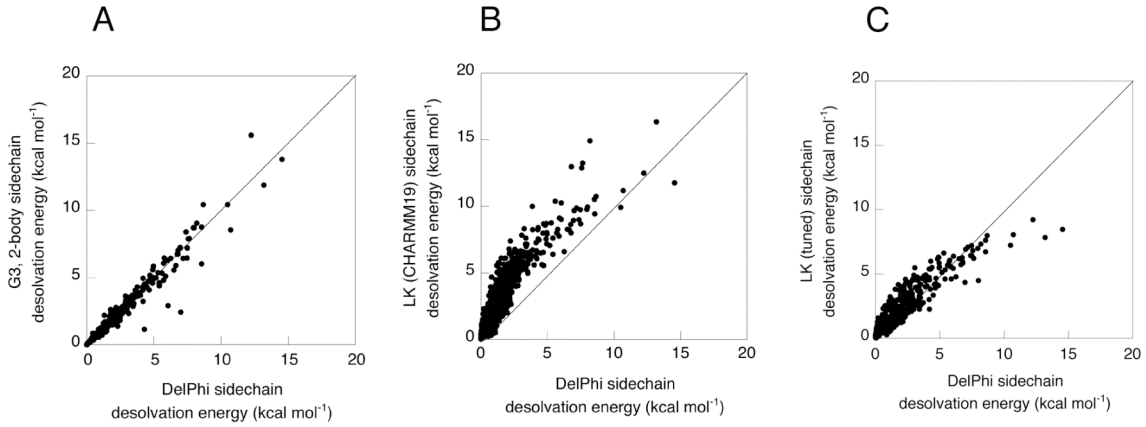


Figure 4-5. Accuracy of the G3 model (A) versus the LK solvent exclusion model (B,C) for approximating sidechain desolvation. Results for both the CHARMM19 (B) and tuned (C) LK parameter sets are shown. All plots contain data for the 758 polar sidechains from the 10 structures listed in Table 4-2 and described in the methods section.

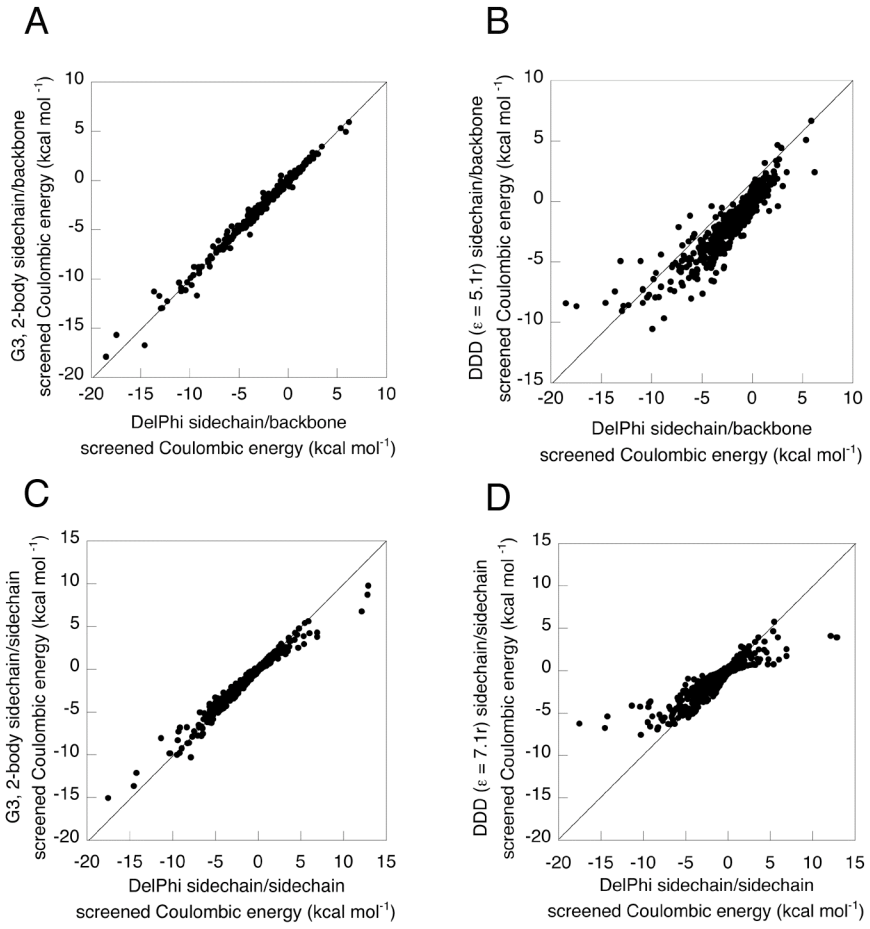


Figure 4-6. Accuracy of the G3 model (A,C) versus the DDD model (B,D) for approximating sidechain/backbone and sidechain/sidechain screened Coulombic interactions. Data is shown for the 19 structures listed in Table 4-2 and described in the methods section.

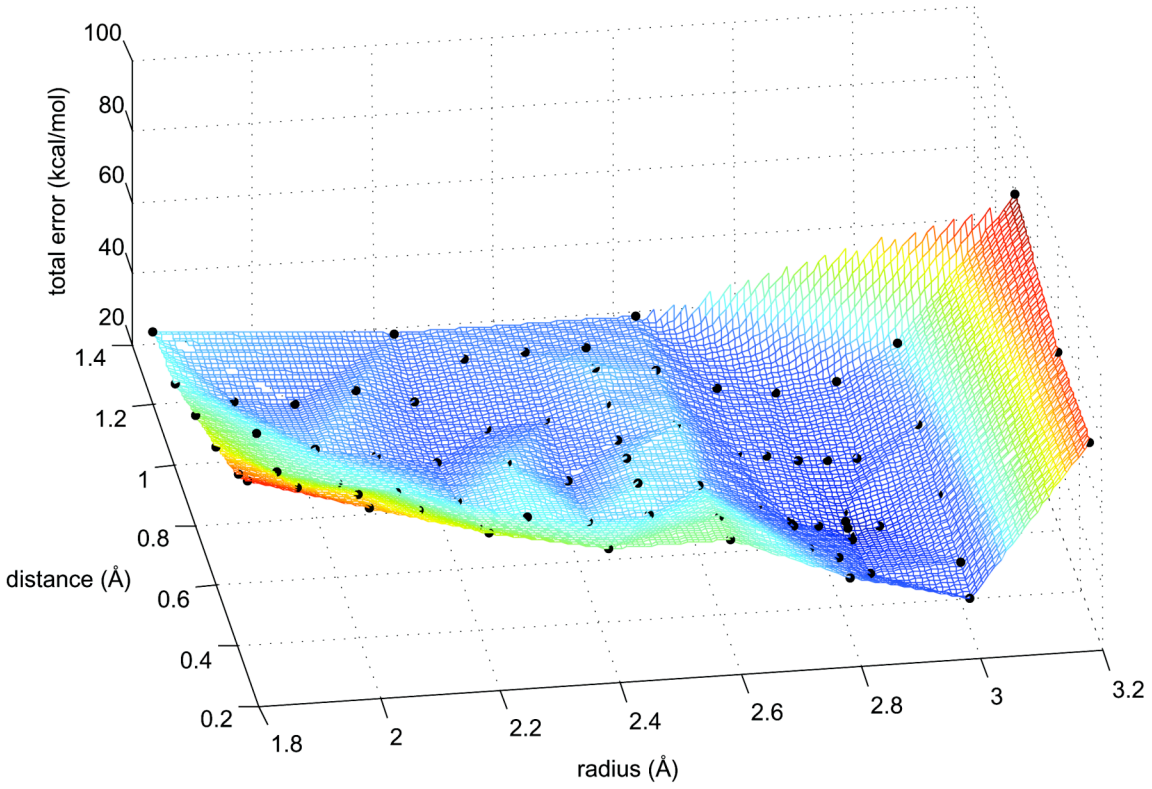


Figure 4-7. The total error associated with varying generic sidechain parameters. The grid surface was obtained by linear interpolation using MATLAB7. The sampled data points are in black.

Chapter 5

Experimental and computational characterization of the Poisson-Boltzmann model in the ORBIT energy function

Abstract

Experimental validation is a crucial test for a protein design energy function. We sought to test the PB model described in Chapter 4 of this thesis by designing the surface residues of *Drosophila melanogaster* engrailed homeodomain (ENH). It had been shown previously that this design target was a valid test case for the electrostatics term in the ORBIT energy function. The PB model does poorly in producing a stabilized variant of ENH. Potential problems with the ENH test case are discussed. In addition, there is a need to accurately model hydrogen bonds in a protein design energy function, and continuum electrostatics model represent an incomplete description of hydrogen bonding. This issue was explored by looking at how the PB model in ORBIT alters the conformations of crystallographic hydrogen bonds. In order to reduce the desolvation penalty associated with the burial of polar atoms, the PB model does not maintain hydrogen-bonding geometries in over half of the cases in a test set of 206 residue pairs. The results described in this chapter provide a starting point for further adaptation of the PB model for protein design.

Introduction

Computational protein design is a tool that provides a starting point for experiments. That starting point is an amino acid sequence that can be synthesized in the lab and characterized for folding, stability, or function. In order to have confidence that a particular design program is valid, the sequences designed by that program must be tested in the laboratory. From a complementary perspective, if computational protein design is a test of our knowledge of the energetics of protein structure, experimental characterization is necessary to provide feed back about our models.[1,2] Other computational benchmarks for protein design have been proposed such as WT sequence recovery,[3] fixed composition design, and point mutation stability prediction.[4] None of these benchmarks is expected to be a sufficient test of a protein design energy function.

We developed a residue pairwise decomposable Poisson Boltzmann (PB) model and computationally tested its ability to reproduce non-pairwise energy calculations.[5] This pairwise model was able to reproduce standard PB calculations better than the solvation and electrostatics terms currently used in the ORBIT energy function.[6] This result indicates that the pairwise PB model is a valid model, worth testing in the ORBIT energy function. The next step is to test the utility of the energy function with the PB model by experimental characterizing sequences designed by that energy function.

Drosophila melanogaster engrailed homeodomain (ENH) is a model system for folding, DNA recognition, and stability.[7-10] Residues 5–56 of this protein (here numbered 1–51), form a well-folded three helix bundle shown in Figure 5-1A.[11] This fragment of the WT protein has an unfolding temperature of 50°C, and a net charge of +5.[12] It has been the subject of many design studies.[12-15] Marshall *et al*. showed that the

stability could be dramatically increased by optimizing the electrostatic properties of the surface residues on ENH.[12] Zollars *et al*. used the same design target to show that sequence design of the surface residues using parameterized dielectrics could also increase the stability of ENH.[14] Thus the surface residues of ENH have become a experimental test of the electrostatics term in the ORBIT energy function. This stability target was chosen instead of an enzyme active site for testing the PB energy function due to the need for and subsequent development of conformational sampling algorithms for placing ligands in active sites.[16] It is also preferable to test energy functions on design targets for which computational design has worked in the past, providing some assurance that other factors in the design process will not confound the results.

In addition to experimental validation using the ENH test case, I was interested in the problem of a consistent treatment of hydrogen bonds in continuum solvation models. Specifically, the continuum dielectric description does not take into account the partial covalent character and orientational constraints of hydrogen bonds.[17] It is expected that that any complete description of protein and active site energetics will require an accurate treatment of hydrogen bonds. Currently, ORBIT and other design programs use an angle-dependent hydrogen-bond term (Fig. 5-6A).[17-19] Since the PB model has not been implemented in a rotamer-based sequence selection scheme like ORBIT, it is unclear whether the PB model would inherently capture any of the geometric properties of ideal hydrogen bonds. Here I investigate this issue by identifying hydrogen-bonded pairs in a set of crystal structures, optimizing the conformation of the sidechains in each pair using various energy functions (including one with the PB model), and finally scoring whether the optimized conformation is hydrogen-bonded. This chapter highlights the current

challenges for using the PB energy function in protein design, starting with experimental validation.

Methods

Experimental

All ENH variants were expressed in a modified pet11 vector with a His_6-Ubiquitin tag N-terminal to the ENH gene. *E. coli* strain BL21-DE3 containing the vector was induced at $OD^{600\ nm}$ = 0.6–0.9 with 1 mM IPTG for 5 hours at 37°C. Cells were harvested by spinning cultures for 15 min at 5000 g. Pellets were resuspended in 20 mM Tris pH 7.4, 10 mM imidazole, 150 mM NaCl, and lysed by french press. Lysates were clarified by centrifuging cultures for 30 minutes at 15000 g. Clarified lysates were batch bound with 3–4 mL Ni-NTA resin for 1 hr at 4°C. Resin, with bound protein was washed with at least 10 column volumes of 30 mM imidazole in 20 mM Tris pH 7.4, 150 mM NaCl. Protein was eluted in 250 mM imidazole. The buffer was changed to 50 mM Tris pH 7.8 and the protein was concentrated to 1–1.5 mL. 1 μL of β-mercaptoethanol was added to the protein sample. 300 μL 1.4 mg/mL His_6-UCH-L3 ubiquitin hydrolase was incubated for 15 minutes at room temperature with 3 μL 1 M DTT. The hydrolase and ENH solutions were mixed gently and incubated for 4–8 hours (without mixing) at 37°C. The reaction solution was batch bound with 3 mL Ni-NTA resin for 1 hr at 4°C and filtered in a gravity column. The flow through was collected and concentrated in 50 mM sodium phosphate at pH 5.5.

Circular dicroism (CD) experiments were carried out on an Aviv 62DS spectrometer. Thermal denaturation data was collected by monitoring the CD signal at

222 nm while the temperature was raised in 1°C steps with a 2 minute equilibration time and 30 second signal averaging time at each step. Protein concentrations were obtained using theoretical extinction coefficients at 280 nm calculated by the program Central Dogma (Huang and Dirks, unpublished). Data was fit to a two-state unfolding transition assuming a temperature-independent heat capacity change ΔC_p.[20] The inflection point of unfolding curves was taken as the temperature *T* at which the value of *signal(T) - signal(T-1)* was maximal.

The PB model in ORBIT

The version of DelPhi used in the calculations in Chapters 3 and 4 of this thesis was implemented into the SETUP module of ORBIT. The default parameters were set to 0.5 Å grid spacing, 70% fill of the grid, 1.4 Å probe radius, and 50 mM ionic strength. Rotamer pairs separated by more than 8.0 Å (minimum distance between polar atoms) were not treated with PB calculations. Their two-body perturbations to each other's desolvation and roatmer/template screened Coulombic energies were ignored, and their rotamer/rotamer screened Coulombic energies were calculated using a distance dependent dielectric of 7.1r. Three-sphere generic sidechains (radius = 2.85 Å and distance from C_α and between spheres = 0.61 Å) were included at positions of unknown identity.

To remove non-physical interactions observed between a rotamer at position *n* and the backbone of position *n±1* in initial designs using the PB model, we used a reference state for interactions between the rotamers and the backbone (Figure 5-2B). Specifically, the rotamer/backbone screened Coulombic energy in the standard truncated

tri-peptide reference state used in ORBIT was subtracted from the one-body rotamer/backbone screened Coulombic energy. These corrected one-body energies, along with the one-body desolvation energies, were stored following the singles energy calculation and used to calculate the two-body perturbation energies to be stored in the rotamer pairs matrix.

ENH designs

Using the same minimized coordinates for ENH used previously,[12] the surface residues of ENH (2, 4, 5, 6, 8, 9, 12, 13, 16, 17, 18, 20, 22, 23, 24, 27, 28, 31, 32, 36, 37, 38, 41, 42, 45, 46, 48, 49, 50) were designed allowing for the amino acids: Asp, Asn, Glu, Gln, His, Lys, Ser, Thr, Ala, and Arg. Histidine was modeled in its positively charged form. The 1996 or 2002 Dunbrack rotamer libraries with no χ expansions were used.[21] For all DelPhi calculations, PARSE[22] atomic radii and charges were used with the parameters described in the previous section. DREIDING atomic radii scaled by 0.9 were used for the Lennard-Jones van der Waals term in ORBIT. For the NC0 and NC3 designs, the energy function included: the scaled van der Waals term, the hydrogen-bonding term with a well depth of 8.0 kcal mol^{-1}, and a Coulombic term with a distance dependent dielectric of 40r. For the NC3 design, residues 4, 22, and 36 were constrained to high propensity N-capping amino acids (Asn, Asp, Ser, Thr); residues 5, 6, 23, 24, 37, and 38 were defined as N-terminal helical residues and were constrained to exclude positively charged amino acids; and residues 16, 17, 31, 32, 49, 50 were defined as C-terminal helical residues and were constrained to exclude negatively charged amino acids. For all designs, the optimal rotameric sequence was obtained using the FASTER

algorithm. The same optimal sequence was found on the majority of the 16 nodes for each design. A list of top-ranking sequences for each design was compiled from 10^8 steps of Monte Carlo sampling, with temperature cycling between 150K and 4000K, starting from the optimal sequence.

Hydrogen-bonding test

In a test set of 18 structures (1igd, 1msi, 1opd, 1fna, 2acy, 1erv, 1dhn, 1whi, 3chy, 2rn2, 1hka, 3lzm, 1amm, and 2pth; plus chain A of: 1agj, 1mol, and 1elk), hydrogen bonds were identified as non-zero interaction energies using the ORBIT angle-dependent hydrogen bond term (Figure 5-5A). Rotamers for the WT amino acids were selected from the backbone-dependent 2002 Dunbrack rotamer library with an expansion of one standard deviation about χ^1 and χ^2 angles.[21] The crystallographic rotamer was also included in the set of available rotamers. The energy function included a van der Waals term (with DREIDING atomic radii scaled by 0.9) plus the term of interest (PB model, distance-dependent dielectric, ORBIT hydrogen-bonding term with damped electrostatics, or nothing). The flow of the computational experiment is shown in Figure 5-5B. The ORBIT hydrogen-bonding term used to select rotamers (Table 5-1) is the same as that which was used to identify hydrogen bonds and rescore energies of new rotamers.

Results

ENH designs

Using the same set of surface residues as in previous studies,[12] the PB model in ORBIT was used optimize the sequence of ENH. In order to compare to previously reported data, we also designed sequences using a standard version of the ORBIT energy function ("NC0") and that same energy function plus sequence constraints relating to the macro-dipole of the α-helix and the N-capping hydrogen bond at the end of the helix ("NC3"). Stability measurements for sequences designed using these three strategies are shown in Figure 5-3B. Using thermostability as the metric, the PB energy function performs the worst of the three methods tested. Since the list of high-scoring sequences for each design contained mutations at positions 8 and 22, residues that were suspected to be important for stability, I made a sampling of sequences from each list in Figure 5-3A. The stabilities of the mutants (light lines in Fig. 5-3B) cluster near the stability of the optimal sequence (dark lines in Fig. 5-3B) for each design strategy.

It is difficult to identify a single factor that leads to the low stability of the PB sequence. There are 14 mutations between the top-scoring NC0 and NC3 sequences, 20 mutations between NC0 and PB, and 17 mutations between NC3 and PB. All variants are over 20-fold mutants of WT. In the top-ranked NC0 designed sequence, there are five violations to the helix dipole rules that were strictly enforced for the NC3 design. There is only one such violation in the PB sequence, Arg6, which does not make any specific interactions with other sidechains or the backbone. Because the statistically derived propensities[21] for the rotamers chosen in the PB design were lower than for those chosen in the other designs, I made a design that took into account those propensities.

This sequence had 16 mutations from the PB design in Figure 5-3, and its stability was higher (T_m = 67°C vs. 50°C) but did not approach the maximal stabilities observed for this design target (> 85°C).

Zollars *et al.* observed a significant increase in stability using the NC0 design strategy when using the 1996 Dunbrack rotamer library instead of the older rotamer library used by Marshall *et al*.[12] Based on this result, I designed and characterized NC3 and PB variants using the 1996 rotamer library (Figure 5-4). It is notable that using this rotamer library, the three design strategies perform almost exactly the same. This points to the problem of how sensitive sequence design is to the parameters unrelated to the energy function (see conclusions section).

Hydrogen bond test

I tested the consistency of the rotameric interactions predicted to be optimal by the PB model in the ORBIT energy function with the hydrogen-bonding term currently used in ORBIT (Figure 5-5A). An angle-dependent hydrogen-bond term has been shown to improve the recovery of WT amino acid identity by the ROSETTA energy function[19] and is expected to be a necessary component of active site design. In a purely electrostatic interaction scheme, the most favorable conformation for a dipole-dipole interaction is linear (Figure 5-5A). For hydrogen bonds, this means 180° angles formed by the donor-hydrogen-acceptor atoms and by the hydrogen-acceptor-acceptor base atoms. This orientation is not observed in protein crystal structures and is not supported by more detailed electronic structure calculations of hydrogen-bonding partners.[17] The PB model includes protein-solvent interactions as well as electrostatic interactions within

the protein, and it is unclear how the balance of these terms might affect the selection of rotamers.

The outline of the computational experiment is shown in Figure 5-5B. For each pair of residues in a set of crystallographic hydrogen bonds, various energy functions were used to select optimized rotamers. Using the ORBIT hydrogen-bonding potential, the hydrogen-bond energies for the rotamers selected by each energy function were evaluated. The results for the various energy functions tested are given in Figure 5-6. The rescored energies of the rotamers selected by the various energy functions were compared to the energies of the crystallographic rotamers to get the value $\Delta E_{h\text{-}bond}$. The sign of $\Delta E_{h\text{-}bond}$ indicates whether the hydrogen-bond energy of the new rotamers is more favorable ($\Delta E_{h\text{-}bond} < 0$), less favorable ($\Delta E_{h\text{-}bond} > 0$), or the same as the crystal structure (Figure 5-6). I also looked at $E_{h\text{-}bond}$ for the new rotamers to see how many pairs maintained any hydrogen-bond character. In order to control for bias introduced by the inclusion of the crystallographic rotamer, the test was carried out using no electrostatics or hydrogen-bonding term ("vdw alone" in Table 5-1). For the 206 pairs examined here, the choice of rotamers by the PB model energy function reduced the magnitude of the hydrogen-bond energy ($\Delta E_{h\text{-}bond} > 0$) for 107 pairs. The distance-dependent dielectric model performed similarly, selecting rotamers with $\Delta E_{h\text{-}bond} > 0$ for 113 pairs. Both of these models are between the two extreme cases: (1) no electrostatic or hydrogen-bonding term (vdw alone), which led to 186 pairs with $\Delta E_{h\text{-}bond} > 0$ and (2) an explicit hydrogen-bond term with damped electrostatics, which gave only 33 pairs with $\Delta E_{h\text{-}bond} > 0$.

To see what factor is driving selection of rotamers in the PB model, for each of the 206 pairs in the test set, the crystallographic rotamers were scored using the PB model and then compared to the PB model energy of the selected rotamers. Figure 5-7 shows each term in the energy function separately. The rotamer desolvation energy is consistently more favorable in the selected rotamers than in the crystallographic rotamers. On the other hand, the screened Coulombic energy (which includes interactions between rotamers as well as interactions between rotamers and the rest of the protein) is generally more favorable in the crystallographic conformation. This result indicates that minimizing the desolvation of rotamers is driving selection of conformations that are not hydrogen bonded.

Conclusions

The data presented here highlight several challenges to implementing the PB model in protein design calculations. The results for ENH leave open the question of experimental validation of the model. Based on the widely varying results with different rotamer libraries, not only for the PB model but also for the NC0 and NC3 designs, I have little confidence in the ENH test case. The sensitivity of the calculation to parameters unrelated to energy function point to the fact energy functions should not be validated or discarded based on experimental characterization of a single lowest energy sequence. Even though there was clustering of stabilities in the very small sequence "libraries" in Figure 5-3, the distribution of stabilities is wide enough to believe that the top 100 or 1000 sequences might show significantly different expectation values for stability from the handful investigated here. In Chapter 7 of this thesis, methods for

computationally designing combinatorial libraries of proteins are evaluated. These methods could be used to carry out high throughput energy function validation in cases where a suitable experimental screen is available.

The other issue addressed here is whether or not the PB model can be considered a replacement for an angle-dependent hydrogen-bond term. Based on the data in Figure 5-6, the PB model does not have any inherent features that would lead it to select hydrogen-bonded rotamers more frequently than a DDD model. It is possible that the force fields used to refine the crystal structure may lead to hydrogen-bond geometries that are non-physical and that certain members of the test set used here are pairs in incorrect conformations. Despite that uncertainty, the results here indicate that the definition of a favorable hydrogen bond in the current ORBIT energy function is not consistent with the geometries produced by the PB model in ORBIT and that to recover these interactions, one would need to account for the angle dependence and covalent character of hydrogen bonds. The more general issue may be energy function balancing: an improvement in one term requires balancing with the other terms. Solving this multiple dimension optimization problem is not straightforward, but there are some computational tests that could serve as a first pass before experimental validation.[3,16,23]

Additional considerations that were not addressed here include computational efficiency of the PB model and unfolded state modeling. The computational efficiency of the PB model as implemented in ORBIT currently would preclude large design problems. This is because for PB designs, computation time scales not only as the square of the number of rotamers but also as the cube of the longest dimension of the molecule, the latter factor due to the grid-based PB solver. Unfolded state modeling is a more general

problem in biomolecular simulation. There are a variety of models in the literature for how one might approach this problem.[24-26] However, in a recent study of several models, they all had the same behavior in reproducing pKa values in an unfolded staphylococcal nuclease,[27] underscoring the difficulty of benchmarking models.

Despite the challenges described in this chapter, much progress has been made toward implementing improved models for electrostatics in protein design. The PB model in ORBIT represents a first step toward developing an environmentally sensitive electrostatics model that includes a consistent treatment of desolvation and electrostatic interactions. Combined efforts in improving efficiency of the numerical solver, experimental validation, and clever modeling will accelerate the pace of functional protein design.

Acknowledgements

I am grateful to Ben Allen for useful discussions about the computational test for hydrogen bonds, Fred Tan for initial help in protein purification, and Jennifer Keeffe for ubiquitin hydrolase stocks and advice on protein purification.

References

1. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.

2. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513.

3. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10383–10388.

4. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.

5. Marshall, S. A., Vizcarra, C. L. & Mayo, S. L. (2005). One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein Sci.* **14**, 1293–1304.

6. Vizcarra, C. L., Zhang, N., Marshall, S. A., Wingreen, N. S., Zeng, C. & Mayo, S. L. (2008). An improved pairwise decomposable finite-difference Poisson-Boltzmann method for computational protein design. *J. Comput. Chem.* **29**, 1153–1162.

7. Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. (2000). Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13518–13522.

8. Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M., Alonso, D. O., Daggett, V. & Fersht, A. R. (2003). The complete folding pathway of a protein from nanoseconds to microseconds. *Nature* **421**, 863–867.

9. Simon, M. D., Feldman, M. E., Rauh, D., Maris, A. E., Wemmer, D. E. & Shokat, K. M. (2006). Structure and properties of a re-engineered homeodomain protein-DNA interface. *ACS Chem. Biol.* **1**, 755–760.

10. Stollar, E. J., Mayor, U., Lovell, S. C., Federici, L., Freund, S. M., Fersht, A. R. & Luisi, B. F. (2003). Crystal structures of engrailed homeodomain mutants: implications for stability and dynamics. *J. Biol. Chem.* **278**, 43699–43708.

11. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L. & Pabo, C. O. (1994). Structural studies of the engrailed homeodomain. *Protein Sci.* **3**, 1779–1787.

12. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* **316**, 189–199.

13. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* **305**, 619–631.

14. Zollars, E. S., Marshall, S. A. & Mayo, S. L. (2006). Simple electrostatic model improves designed protein sequences. *Protein Sci.* **15**, 2014–2018.

15. Shah, P. S., Hom, G. K., Ross, S. A., Lassila, J. K., Crowhurst, K. A. & Mayo, S. L. (2007). Full-sequence computational design and solution structure of a thermostable protein variant. *J. Mol. Biol.* **372**, 1–6.

16. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. (2006). Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16710–16715.

17. Morozov, A. V. & Kortemme, T. (2005). Potential functions for hydrogen bonds in protein structure prediction and design. *Adv. Prot. Chem.* **72**, 1–38.

18. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.

19. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239–1259.

20. Minor, D. L., Jr. & Kim, P. S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. *Nature* **367**, 660–663.

21. Dunbrack, R. L., Jr. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.

22. Sitkoff, D., Sharp, K. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978–1988.

23. Hom, G. K. & Mayo, S. L. (2006). A search algorithm for fixed-composition protein design. *J. Comput. Chem.* **27**, 375–378.

24. Creamer, T. P., Srinivasan, R. & Rose, G. D. (1995). Modeling unfolded states of peptides and proteins. *Biochem.* **34**, 16245–16250.

25. Fitzkee, N. C. & Rose, G. D. (2004). Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12497–12502.

26. Jha, A. K., Colubri, A., Freed, K. F. & Sosnick, T. R. (2005). Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13099–13104.

27. Fitzkee, N. C. & Garcia-Moreno, E. B. (2008). Electrostatic effects in unfolded staphylococcal nuclease. *Protein Sci.* **17**, 216–227.

Table 5-1. Change in hydrogen-bond energy

	PB	*DDD*[*]	*ORBIT h-bond*[†]	*vdw alone*[‡]
$E_{h\text{-}bond} < 0$[§]	157	159	186	64
$\Delta E_{h\text{-}bond} \leq 0$	99	93	173	20
$\Delta E_{h\text{-}bond} \geq 0$	107	113	33	186

[*] DDD model used ε = 7.1r for sidechain/sidechain interactions and ε = 5.1r for sidechain/backbone interactions.
[†] ORBIT angle-dependent hydrogen-bonding term with a well-depth of 8.0 kcal mol^{-1} and ε = 40r
[‡] No other energy terms other than the van der Waals potential were used.
[§] The number of pairs that retain any hydrogen bond character after rotamer selection

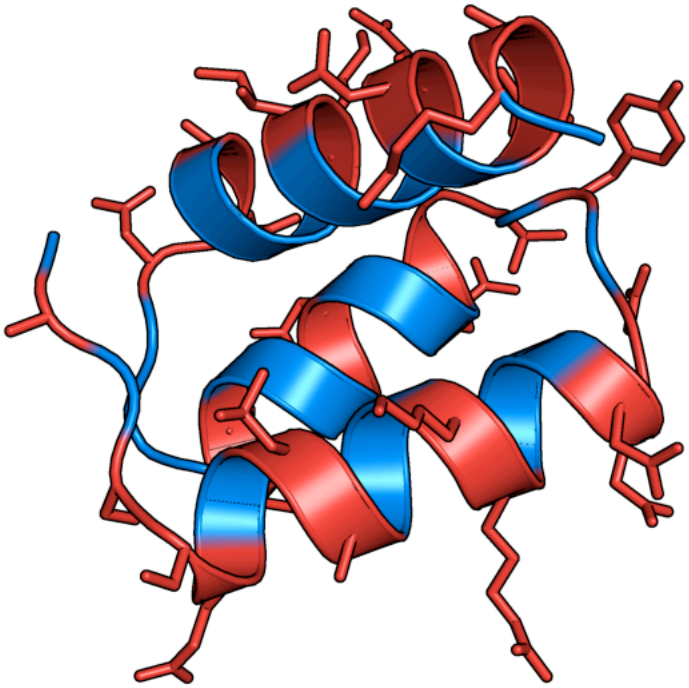


Figure 5-1. The ENH test case. The surface residues that were designed in this study are shown in red, with side chains shown as sticks.

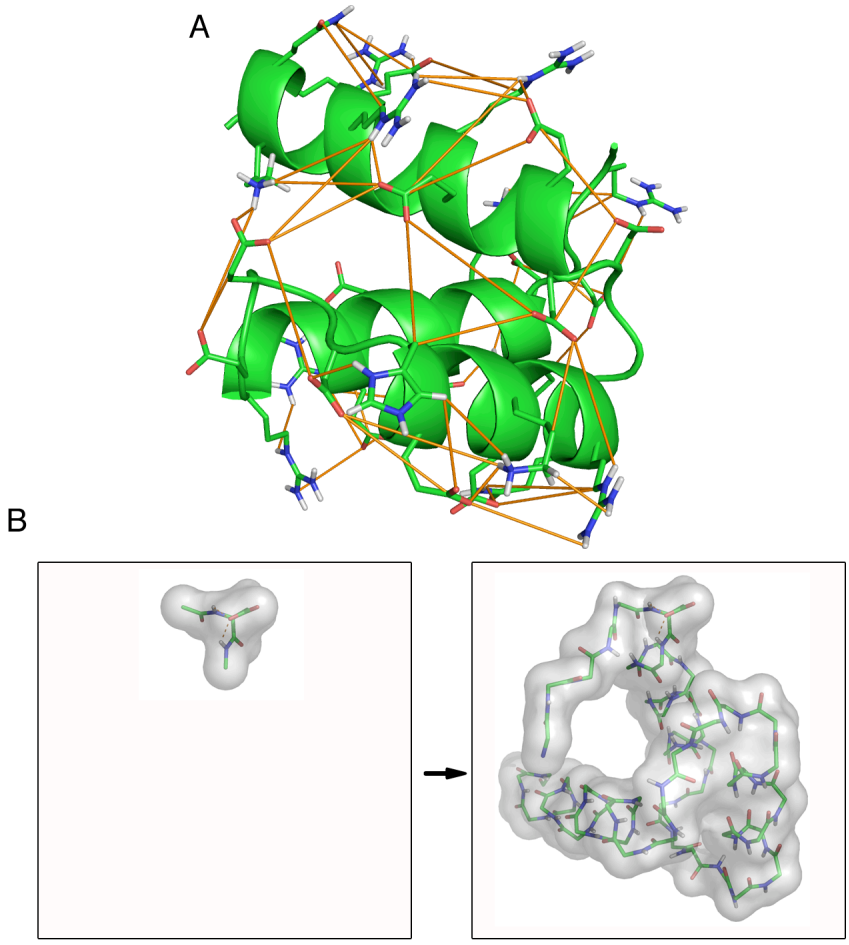


Figure 5-2. The PB model in ORBIT. (A) Illustration of the cutoffs used for interactions between sidechains. All of the rotamer/rotamer interactions that were treated with PB electrostatics in the PB design are connected by orange lines. Only polar hydrogens are shown. Molecule is rotated 180° about the vertical axis from the image in Figure 5-1. (B) Reference state for sidechain/backbone interactions. For each rotamer, the rotamer/backbone screened Coulombic energy in the truncated tri-peptide reference state (left) used in ORBIT was subtracted from the one-body folded state rotamer/backbone screened Coulombic energy (right).

A

NC0:

		2	4	5	6	8	9	12	13	16	17	18	20	22	23	24	27	28	31	32	36	37	38	41	42	45	46	48	49	50
1	-156.4196	ASP	SER	ASP	ARG	GLU	GLU	ARG	LYS	GLU	GLU	ASN	GLU	SER	GLU	LYS	ASN	GLU	LYS	GLU	ASP	GLN	GLU	GLU	ARG	GLN	GLU	LYS	ARG	GLN
2	-156.3758	|	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
3	-156.3742	ASN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
4	-156.3389	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU
5	-156.3304	ASN	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
6	-156.3145	ASN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU
7	-156.3118	ARG	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
8	-156.2951	|	|	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU
9	-156.2728	ARG	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU
10	-156.2707	ASN	|	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU

NC3:

		2	4	5	6	8	9	12	13	16	17	18	20	22	23	24	27	28	31	32	36	37	38	41	42	45	46	48	49	50
1	-149.1635	ASP	SER	ASP	GLN	GLU	ARG	ASP	GLU	ARG	ARG	ASN	GLU	SER	ASP	GLN	ARG	GLU	HSP	ARG	ASP	GLU	GLU	GLU	ARG	GLN	GLU	LYS	ARG	GLN
2	-149.1162	|	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
3	-149.0994	|	|	|	|	LYS	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
4	-149.0846	ASN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
5	-149.0521	|	|	|	|	LYS	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
6	-149.0373	ASN	|	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
7	-149.0181	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	ARG	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
8	-148.9809	ARG	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
9	-148.9746	|	|	|	|	|	|	|	|	|	|	ASP	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
10	-148.9709	|	|	|	|	|	|	|	|	|	|	|	|	|	THR	|	|	ARG	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
11	-148.9599	|	|	|	|	LYS	|	|	|	|	|	|	|	|	|	|	ARG	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
12	-148.9576	|	|	|	|	LYS	|	|	|	|	|	ASP	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|

PB:

		2	4	5	6	8	9	12	13	16	17	18	20	22	23	24	27	28	31	32	36	37	38	41	42	45	46	48	49	50
1	-139.2449	ARG	ASP	GLU	ARG	LYS	GLU	ASP	GLU	ARG	ARG	GLU	GLU	ASP	HSP	GLU	LYS	ASP	ARG	GLN	ASP	GLU	GLU	GLU	ARG	ARG	GLU	LYS	GLN	ARG
2	-139.1917	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	HSP	|	|	|	|	|	|	|	|	|	|	|	|	|	|
3	-139.1620	|	|	|	|	|	|	|	|	|	|	|	ASN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU	|
4	-139.1194	|	|	|	|	|	|	|	|	|	|	|	ASP	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
5	-139.0848	|	|	|	|	|	|	|	|	|	|	|	ASN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
6	-139.0814	|	|	|	|	GLN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
7	-139.0651	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	GLU	|
8	-139.0562	|	|	|	|	|	|	|	ASP	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
9	-139.0442	|	|	|	|	|	|	|	|	|	|	|	|	LYS	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
10	-139.0429	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	HSP
11	-139.0423	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	LYS	|
12	-139.0278	|	|	|	|	|	|	|	|	|	|	|	ASN	|	|	|	HSP	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
13	-138.9962	|	|	|	|	GLN	|	|	|	|	|	|	ASN	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|

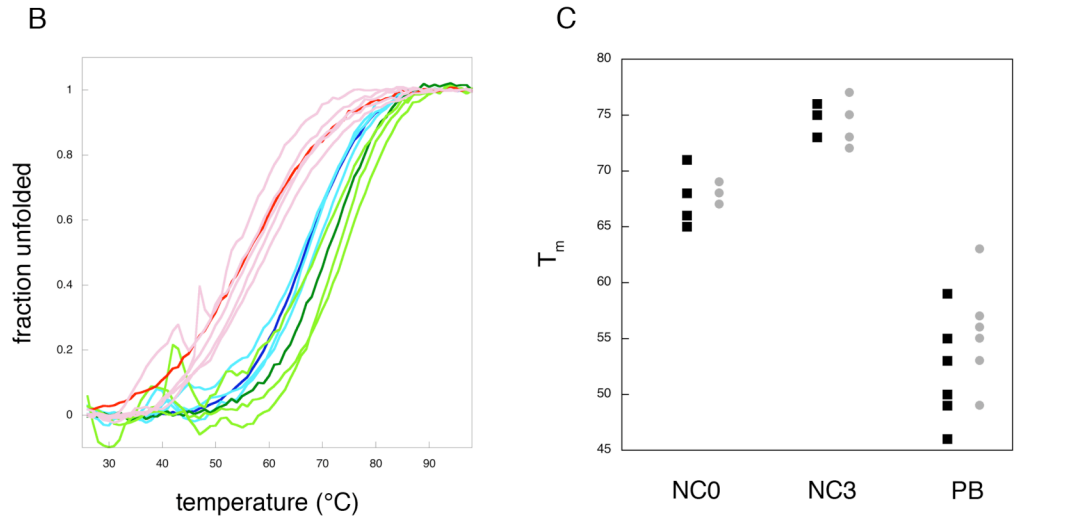


Figure 5-3. Experimental data for the ENH variants. (A) List of low energy sequences for each design. (B) Thermal denaturations monitored by CD. The dark line is for the lowest energy sequence for each design, sequence in (A). The lighter lines correspond to high-scoring sequences found by Monte Carlo sampling around the optimal sequence, highlighted in (A). (C) Plot of T_m values calculated using the inflection point (black squares) or two-state non-linear fit (gray circles) of the curves in (B).

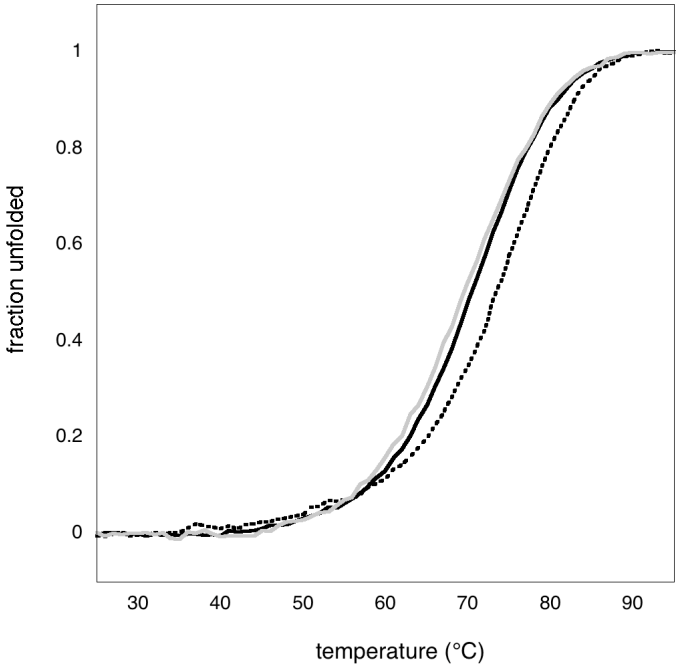


Figure 5-4. CD data for the ENH variants designed using 1996 rotamer library. Thermal denaturations monitored by CD are shown for NC0 (dashed line), NC3 (black line), and PB (gray line) variants. The NC0 data is from Zollars et al.[14]

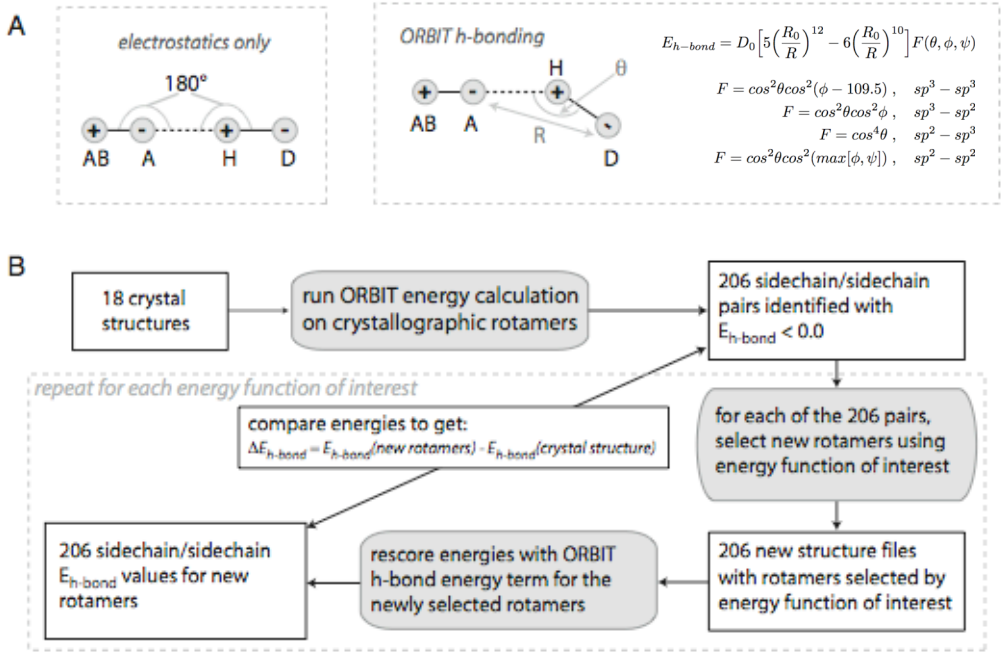


Figure 5-5. Test of hydrogen-bonding geometries. (A) Idealized geometry for a purely electrostatic interaction is shown on the left. The ORBIT hydrogen bond energy term is shown on the right with equations for energy and F for each type of donor-acceptor pair (θ = donor-hydrogen-acceptor angle, ϕ = hydrogen-acceptor-base angle, φ = angle between normals of the planes defines by the six atoms attached to the sp^2 centers, R_0 = 2.8 Å, D_0 = 8.0 kcal mol^{-1}). (B) Scheme for testing the degree to which the PB model selects rotamers that have hydrogen-bonding geometries. The initial identification of crystallographic hydrogen bonds was carried out once. The steps enclosed by the dashed line were done for each energy function listed in Table 5-1.

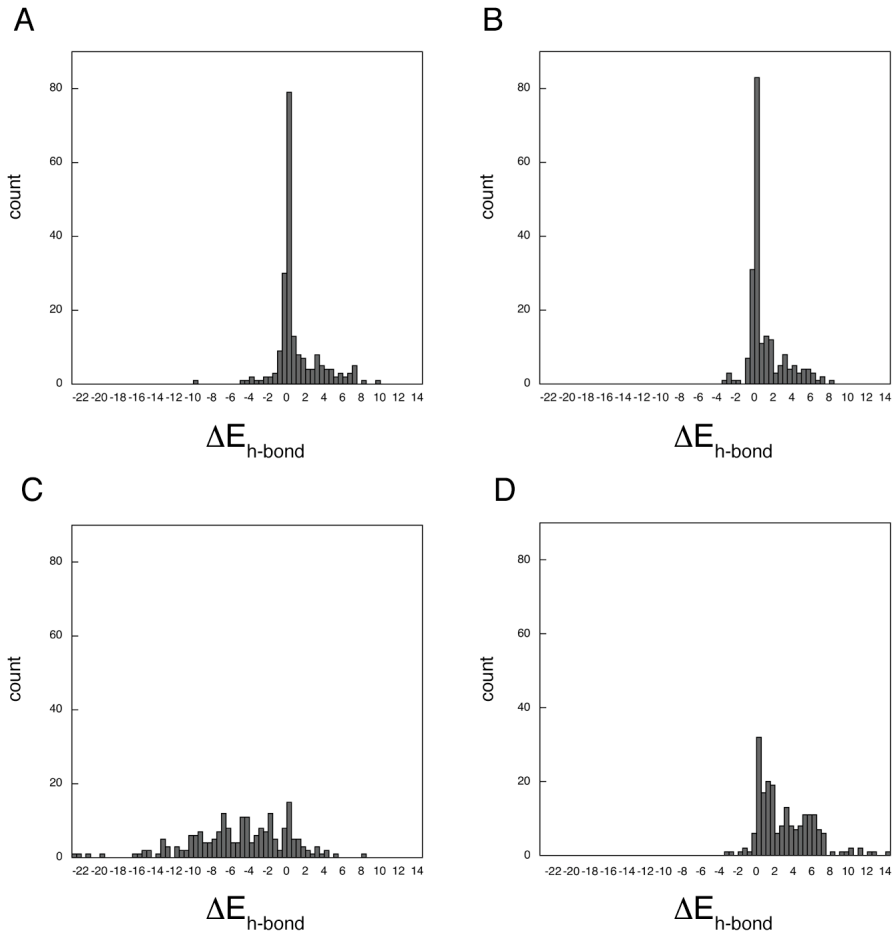


Figure 5-6. Results of the hydrogen bond test. For (A) PB model, (B) DDD model, (C) ORBIT energy function with hydrogen bond potential, and (D) van der Waals potential alone. As described in the text, $\Delta E_{h\text{-}bond} = E_{h\text{-}bond}$*(selected rotamers)* $- E_{h\text{-}bond}$*(crystal structure)*.

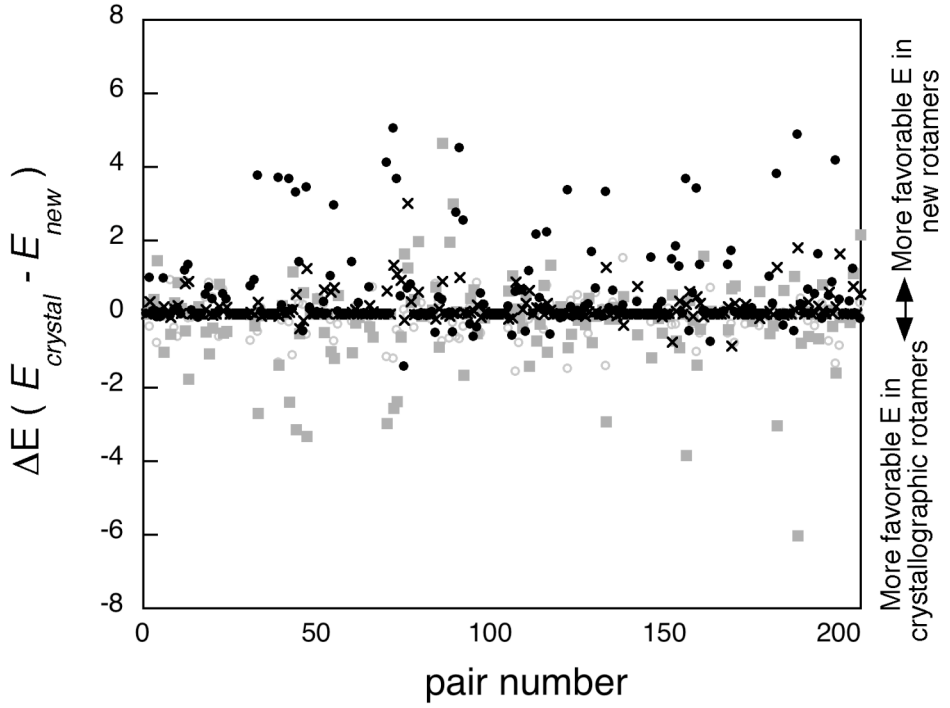


Figure 5-7. Desolvation penalty leads to loss of hydrogen-bonding geometries. For each of the 206 pairs in the test set, the change in each force field term between the new and selected selected rotamer is shown (van der Waals energy, open circles; electrostatic interaction energy, gray squares; polar desolvation of the backbone, black "x"; polar desolvation of sidechains, black circles).

Chapter 6

The plasticity of surface residues on engrailed homeodomain

Abstract

I have created a hyper-thermostable variant of Drosophila melanogaster engrailed homeodomain by mutating its solvent-exposed residues. This variant, HT_ENH, was designed using structure-based computational optimization of the amino acid sequence, with sequence constraints for N-capping and helix dipole rules. HT_ENH has a thermal denaturation temperature approximately ~ 50°C higher than the WT sequence. This result is yet another example of protein stabilization via electrostatic optimization of surface residues. In order to reconcile HT_ENH's high stability with recent reports of well-folded proteins with poor surface electrostatic interactions and large formal charges, I also created a negatively "supercharged" variant of engrailed homeodomain with a formal charge of -13. The super-charged variant has circular dichroism similar to the WT sequence. The thermostability is also close to that of the WT. As would be predicted for electrostatic repulsion or optimization, the super-charged variant is stabilized, while HT_ENH is destabilized, by the addition of monovalent salts. These data indicate that for the engrailed homeodomain fold, the plasticity of surface residues can be exploited to either impart highly unfavorable electrostatic interactions without disrupting the fold or to increase the protein's thermostability by optimizing electrostatic interactions.

Introduction

It is now widely accepted that electrostatic optimization of solvent-exposed residues can impart thermodynamic stability on the folded state of proteins. This stabilization has been observed in both naturally occurring and designed proteins.[1-8] For mesophile/thermophile pairs, there are dramatic examples of electrostatic optimization. Perhaps the most studied is in the bacterial cold shock proteins.[9,10] A single mutation of a surface Glu to an Arg caused at least a 2.5 kcal mol^{-1} increase in stability of a mesophilic cold shock protein.[11] Larger database surveys have shown an increased number of close range electrostatic interactions[12] and optimization of longer-range electrostatic properties[13] in thermophilic homologues.

There are also a number of cases of rational design of surface residues leading to stabilized variants. Makhatadze and coworkers have taken a minimalist approach the engineering surface electrostatics by identifying a handful of residues that are likely to have a larger impact on the stability of the protein. Mutagenesis of these residues is predicted to lead to a more optimal charge distribution, and this result as been confirmed in the lab by experimentally characterizing the stability of their rational mutations in a number of different folds.[7,8] Mayo and coworkers have optimized the surface residues of engrailed homeodomain by taking into account the properties of the helix macrodipole and helical N-capping interactions or using a Coulombic electrostatics term, creating variants with melting temperatures at least 35°C above the WT protein.[3,6]

On the other hand, recently it has been shown that folded proteins can tolerate a high net charge and a theoretically large amount of electrostatic repulsion between surface sidechains. Both by chemical modification and mutagenesis, variants of proteins

that naturally have modest net charge, have been created that have extreme charges (±20). Liu and coworkers created a series of GFP variants that had at least ±30 net charges by mutating surface residues.[14] These "supercharged" variants showed an increased resistance to aggregation compared to the parent GFP from which they were derived. Whitesides and coworkers have used chemical modification of surface lysines on bovine carbonic anhydrase to study the role of surface charge in stability, folding kinetics, and ligand binding.[15-17] The modified enzymes have net charges on the order of -19 and retain the ability to refold from chemical denaturants.[17] In addition, removal of all charge/charge interactions on the surface of ubiquitin did not significantly reduce the stability from that of the WT protein, indicating that favorable electrostatic interactions are not a prerequisite for folding or stability.[18]

Taken together, these studies produce a conflicting picture of the role of surface electrostatics in protein stability. It is clear that favorable electrostatic interactions are neither sufficient nor necessary for producing a compact, well-folded protein.[14,17,18] In cases where specific residue pairs were shown to have a small free energy of interaction, it was proposed that these interactions could aid in folding and binding specificity.[19] It is also proposed that proteins might have charged sidechains on their surface in order to prevent aggregation but carry low net charges in order to prevent non-specific binding to oppositely charged molecules.[20]

However, it is also clear that, in some cases, altering the electrostatic properties of protein surfaces can increase the thermodynamic stability of the folded state.[7,8,11] One way to reconcile these somewhat contradictory ideas it to think of stability as a threshold.[21] It is possible that rarely is the electrostatic free energy of folding associated

with the surface residues so unfavorable as to cross that threshold and favor the denatured state. But it is possible that optimization of the electrostatic free energy can be used to increase that the stability further above that threshold. A possible physical explanation for this qualitative model is that polar solvent and mobile counter ions can almost always compensate for unfavorable surface electrostatic interactions between surface residues on a protein.

In this chapter, data is presented for two variants of *Drosophila melanogaster* engrailed homeodomain (ENH). One variant, HT_ENH ("hyperthermophilic ENH"), is highly stabilized compared to the WT protein. It was designed using a similar strategy as previous stabilized sequences, but HT_ENH retained certain WT contacts. In order to investigate the paradox of protein surface charge and stability, I created a negatively "supercharged" variant (NSC) that has 20 of its 29 surface residues mutated to acidic amino acids. Together these variants show the plasticity of surface residues in ENH to stabilize the native fold or to carry extreme charges.

Methods

Experimental methods. All proteins were expressed and purified as described in Chapter 5. Circular dichroism (CD) data was collected on an Aviv 62DS spectrometer. Unless otherwise noted, conditions were 50 mM sodium phosphate buffer at pH 6.5. Sedimentation velocity experiments were done on a Beckman XL-I Ultima analytical ultracentrifuge equipped with absorbance optics. Samples were spun at 55,000 rpm (20°C) at protein concentrations of 75 uM and 150 uM in 50 mM sodium phosphate (pH 6.5) and 150 mM NaCl with detection at 280 nm every 3 minutes. SEDNTERP was used

to calculate the buffer density (1.00886 g mL^{-1}) and partial specific volumes (0.7306 mL g^{-1} for WT ENH and 0.7160 mL g^{-1} for HT_ENH). The data was analyzed using SEDFIT. Both data sets collected for 75 μM protein concentration had RMSD values of 0.005 for their fits (Fig. 6-4). The data for 150 μM HT_ENH had an RMSD of 0.009 associated with its fit, but the absorbance was slightly higher than the optimal range for the instrument.

Computation. Designs were carried out on the 29 surface residues of ENH (2, 4, 5, 6, 8, 9, 12, 13, 16, 17, 18, 20, 22, 23, 24, 27, 28, 31, 32, 36, 37, 38, 41, 42, 45, 46, 48, 49, 50) using the same backbone coordinates as previously.[3,6] These positions were allowed to have the following polar amino acid identities: Asp, Asn, Glu, Gln, His, Lys, Ser, Thr, Ala, and Arg. Histidine was modeled in its positively charged form. For the HT_ENH design, residues 4, 22, and 36 were constrained to high-propensity N-capping amino acids (Asn, Asp, Ser, Thr); residues 5, 6, 23, 24, 37, and 38 were defined as N-terminal helical residues and were not allowed to mutate to positively charged amino acids; and residues 16, 17, 31, 32, 49, 50 were defined as C-terminal helical residues and were not allowed to mutate to negatively charged amino acids. The 2002 Dunbrack rotamer library[22] (with no expansions) plus the crystallographic rotamers were used to model conformational flexibility. Since the crystallographic rotamer was included in the available rotamers, some positions were allowed to remain as their WT non-polar amino acid or a WT amino acid that would violate the helix dipole rules. For the NSC design, Asp and Glu rotamers or the crystallographic rotamer were allowed at each position.

DelPhi was used to calculate electrostatic potential surfaces for the WT crystal structure (pdb code: 1ENH) and for the design models of HT_ENH and NSC (Figure 6-2). PARSE atomic radii and charges were used with a grid spacing of 0.5 Å, probe radius of 1.4 Å, 70% grid fill, $\varepsilon_{water} = 80$, and $\varepsilon_{protein} = 4$. Electrostatic interaction energy for each residue shown in Figure 6-2 was calculated by summing the DelPhi solvent-screened interaction energies between the residue of interest and the other polar sidechains and the backbone.

Results

The HT_ENH sequence was designed by allowing ORBIT to select only amino acids favored according to known properties of the α-helix macrodipole and capping interactions. Specifically, positively charged amino acids were excluded from the three most N-terminal residues of a helix, and negatively charged amino acids were excluded from the three most C-terminal residues of a helix. In addition, the N-capping residues were forced to be high propensity N-capping residues.[23] The helix dipole rules account for long-range electrostatic effects that may be poorly model by the energy function. The ORBIT energy function was used with an angle-dependent hydrogen-bonding term, a damped Coulombic term (ε=40r), and a Lennard-Jones van der Waals term. No solvation model was used on the assumption that all surface residues would be desolvated to the same amount. Unlike previous designs using this approach (see Chapter 5 and Marshall *et al.*[3]), the crystallographic rotamer was included in the set of rotamers considered in the calculation. This was also true of the negatively supercharged (NSC) ENH variant that I designed. Interestingly, in both cases Leu8 was chosen to remain as the WT amino acid.

There is evidence that in solution the sidechain of Leu8 may be more buried than in crystal structures of ENH.[24] Therefore, the classification of this residue as being on the protein surface might be inaccurate.

The models and sequences for the ENH variants are shown in Figure 6-1. The model of HT_ENH shows more short-range (< 4Å), intra-helical sidechain/sidechain interactions than the WT sequence or NSC (Figure 6-1B) but not as many as reported for previous generation designs.[3] Not surprisingly, the sidechains of NSC are not making short-range interactions, and if the design model did show them, it is probable that the sidechains would relax in solution to minimize repulsion. The formal charge listed for NSC in Figure 6-1C is most likely more negative than the actual net charge in solution due to shifted pKa values of acidic groups and bound counterions.[20] Electrostatic potential surfaces for the WT crystal structure and the design models are shown in Figure 6-2. The solvent-screened Coulombic energy between each polar amino acid and the rest of the protein is also shown. The HT_ENH model has almost no unfavorable electrostatic interactions, while WT and NSC have multiple residues with $\Delta G_{electrostatic} >> 0$. It is notable that there are several basic sidechains in the NSC protein that have highly favorable interactions. It is possible that strong favorable interactions with the few remaining oppositely charged residues partially compensate for the large repulsion in highly charged proteins.[14,15]

Stability toward thermal denaturation was measured using circular dichroism (CD) (Figure 6-3). Both HT_ENH and NSC unfold reversibly in native buffer (data not shown). An attempt was made to calculate ΔG of unfolding for HT_ENH. However, HT_ENH does not unfold in 10 M urea, and guanidinium was avoided as denaturant

since it is charged. Therefore, I examined thermal unfolding as an approximation of thermodynamic stability, a sound approximation for a range of proteins.[25] Only an apparent T_m could be obtained for HT_ENH since it does not complete the unfolding transition below 100°C (Figure 6-3B). The inflection point of the unfolding curve is 94°C. Since the high stability of HT_ENH could be due to oligomerization of the protein, I investigated the oligomerization state using analytical ultracentrifugation. On the timescale of a sedimentation velocity run, HT_ENH appears to be monomeric (Fig. 6-4). The unfolding temperature of NSC (48°C) is near that of WT (44°C). The unfolding transition for NSC is more cooperative than that of WT as judged by the slope of the transition region. The denatured state of ENH has been observed to contain significant native secondary structure, leading to a less cooperative folding transition.[26] It is possible that NSC has a more extended denatured state than the WT protein and therefore shows more cooperative unfolding behavior.

The stability of HT_ENH and NSC at varying ionic strength was tested. The T_m of NSC increases monotically with increasing concentration of sodium chloride, while opposite trend was observed for HT_ENH (Figure 6-5). This trend is consistent with unfavorable electrostatic interactions in NSC and favorable interactions in HT_ENH. Pace and coworkers hypothesized that salt screening will be stronger in the unfolded state than in the folded of protein, and therefore, the addition of monovalent salt will differentially affect folded and unfolded state electrostatic interactions.[27] This will only affect the steepness of the slope, not the sign. It seems more plausible that the salt screening of the native state interactions are dominating the different signs of the slopes observed in Figure 6-5. As expected, the stability of NSC is also highly pH dependent

and the thermal dentauration temperature increases from 48°C to 59°C when the pH is lowered from 6.5 to 5.5 (data not shown).

Conclusions

The data here indicate that both HT_ENH and NSC assume the native fold of ENH. Crystals of HT_ENH have been obtained, and structure determination is in progress. These two designed proteins represent two extremes of plasticity for surface residues. If we consider plasticity the ability to change into many distinct forms, HT_ENH represents a surface charge distribution that is electrostatically optimized and NSC represents a charge distribution with a high degree of electrostatic repulsion. Both of these properties were engineered into a fold that has neither of these properties naturally. A survey of the UniProt proteome for *D. melanogaster* showed that NSC had a larger magnitude net charge than any annotated gene with the exception of the L39 ribosomal protein (51 a.a.; +18). It should be noted that the design and expression of supercharged proteins is not trivial. Liu and coworkers report poor expression or aggregation for several of their variants.[14] Efforts to make a positively super-charged ENH have not been successful (data not shown).

In the introduction, I put forward a model for the importance of long-range electrostatic interactions for stabilization of the folded state in which unfavorable interactions can be mitigated by interaction with the highly polarizable solvent and mobile ions. This model might reconcile some of the contradictory results that have been reported in the literature for the importance of surface electrostatics. The molecules reported here are consistent with this idea, but do not count as proof. Proving this model

may be quite difficult. Measurements of sidechain dynamics might shed light on whether the surface sidechains in HT_ENH are less mobile than those in NSC due to the specific favorable interactions they are making with the rest of the protein. There is no reported method for studying the dynamics of non-aliphatic sidechains in solution. Capillary electrophoresis might shed light on hydrodynamic drag and also provide a more direct measure of the net charge of NSC in solution.[20] NMR studies of backbone dynamics may provide information about the heterogeneity of the folded state ensembles of the two variants. Based on the data presented here, these variants represent in extreme of surface electrostatics in the ENH fold. In the future, they could be used to explore the role of surface electrostatics in the folding pathway and folded state dynamics.

Acknowledgements

I am grateful to Possu Huang for useful discussion and help with setting up the analytical ultracentrifugation experiment, Len Thomas and Pavle Nikolovski for crystallography help, and Rich Olson for help with running the analytical ultracentrifuge and data analysis.

References

1. Grimsley, G. R., Shaw, K. L., Fee, L. R., Alston, R. W., Huyghues-Despointes, B. M. P., Thurlkill, R. L., Scholtz, J. M. & Pace, C. N. (1999). Increasing protein stability by altering long-range coulombic interactions. *Protein Sci.* **8**, 1843–1849.

2. Martin, A., Sieber, V. & Schmid, F. X. (2001). In-vitro selection of highly stabilized protein variants with optimized surface. *J. Mol. Biol.* **309**, 717–726.

3. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* **316**, 189–199.

4. Makhatadze, G. I., Loladze, V. V., Ermolenko, D. N., Chen, X. F. & Thomas, S. T. (2003). Contribution of surface salt bridges to protein stability: Guidelines for protein engineering. *J. Mol. Biol.* **327**, 1135–1148.

5. Makhatadze, G. I., Loladze, V. V., Gribenko, A. V. & Lopez, M. M. (2004). Mechanism of thermostabilization in a designed cold shock protein with optimized surface electrostatic interactions. *J. Mol. Biol.* **336**, 929–942.

6. Zollars, E. S., Marshall, S. A. & Mayo, S. L. (2006). Simple electrostatic model improves designed protein sequences. *Protein Sci.* **15**, 2014–2018.

7. Strickler, S. S., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V. & Makhatadze, G. I. (2006). Protein stability and surface electrostatics: A charged relationship. *Biochemistry* **45**, 2761–2766.

8. Schweiker, K. L., Zarrine-Afsar, A., Davidson, A. R. & Makhatadze, G. I. (2007). Computational design of the Fyn SH3 domain with increased stability through optimization of surface charge-charge interactions. *Protein Sci.* **16**, 2694–2702.

9. Perl, D. & Schmid, F. X. (2001). Electrostatic stabilization of a thermophilic cold shock protein. *J. Mol. Biol.* **313**, 343–357.

10. Wunderlich, M., Martin, A. & Schmid, F. X. (2005). Stabilization of the cold shock protein CspB from Bacillus subtilis by evolutionary optimization of Coulombic interactions. *J. Mol. Biol.* **347**, 1063–1076.

11. Perl, D., Mueller, U., Heinemann, U. & Schmid, F. X. (2000). Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat. Struct. Biol.* **7**, 380–383.

12. Szilagyi, A. & Zavodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **8**, 493–504.

13. Torrez, M., Schultehenrich, M. & Livesay, D. R. (2003). Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces. *Biophys J* **85**, 2845–2853.

14. Lawrence, M. S., Phillips, K. J. & Liu, D. R. (2007). Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **129**, 10110–10112.

15. Gudiksen, K. L., Gitlin, I., Yang, J., Urbach, A. R., Moustakas, D. T. & Whitesides, G. M. (2005). Eliminating positively charged lysine epsilon-NH3+ groups on the surface of carbonic anhydrase has no significant influence on its folding from sodium dodecyl sulfate. *J. Am. Chem. Soc.* **127**, 4707–4714.

16. Gitlin, I., Gudiksen, K. L. & Whitesides, G. M. (2006). Peracetylated bovine carbonic anhydrase (BCA-Ac-18) is kinetically more stable than native BCA to sodium dodecyl sulfate. *J. Phys. Chem. B* **110**, 2372–2377.

17. Gitlin, I., Gudiksen, K. L. & Whitesides, G. M. (2006). Effects of surface charge on denaturation of bovine carbonic anhydrase. *Chembiochem* **7**, 1241–1250.

18. Loladze, V. V. & Makhatadze, G. I. (2002). Removal of surface charge-charge interactions from ubiquitin leaves the protein folded and very stable. *Protein Sci.* **11**, 174–177.

19. Strop, P. & Mayo, S. L. (2000). Contribution of surface salt bridges to protein stability. *Biochemistry* **39**, 1251–1255.

20. Gitlin, I., Carbeck, J. D. & Whitesides, G. M. (2006). Why are proteins charged? Networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis. *Angew. Chem., Int. Ed.* **45**, 3022–3060.

21. Bloom, J. D., Arnold, F. H. & Wilke, C. O. (2007). Breaking proteins with mutations: threads and thresholds in evolution. *Mol Syst Biol* **3**, 76.

22. Dunbrack, R. L., Jr. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6**, 1661–1681.

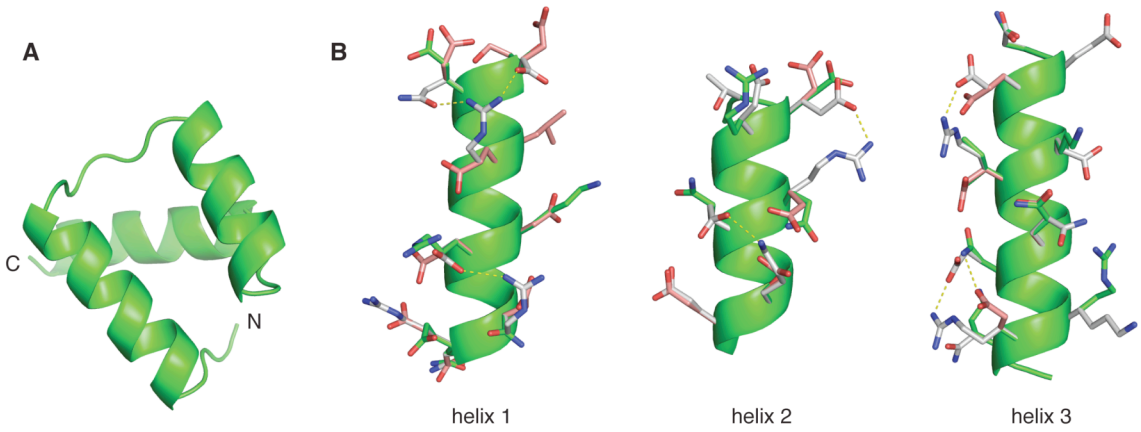
23. Seale, J. W., Srinivasan, R. & Rose, G. D. (1994). Sequence determinants of the capping box, a stabilizing motif at the N-termini of alpha-helices. *Protein Sci* **3**, 1741–1745.

24. Religa, T. L. (2008). Comparison of multiple crystal structures with NMR data for engrailed homeodomain. *J. Biol. NMR* **40**, 189–202.

25. Rees, D. C. & Robertson, A. D. (2001). Some thermodynamic implications for the thermostability of proteins. *Protein Sci.* **10**, 1187–1194.

26. Mayor, U., Grossmann, J. G., Foster, N. W., Freund, S. M. V. & Fersht, A. R. (2003). The denatured state of engrailed homeodomain under denaturing and native conditions. *J. Mol. Biol.* **333**, 977–991.

27. Pace, C. N., Alston, R. W. & Shaw, K. L. (2000). Charge-charge interactions influence the denatured state ensemble and contribute to protein stability. *Protein Sci.* **9**, 1395–1398.



C

```
          1--------10--------20--------30--------40--------50      Q
WT        TAFSSEQLARLKREFNENRYLTERRQQLSSELGLNEAQIKIWFQNKRAKI      +5
HT_ENH    |D|SDQ|LR||DE||RRN|E|TEE||RE||KE|G|DEE||ER||QE|KRQ|      -2
NSC       |D|SEE|LE||DE||EED|Y|TER||EQ||EE|G|DEE||EE||EN|REK|      -13
```

Figure 6-1. Designed ENH variants. (A) Cartoon representation of the ENH structure (pdb code: 1ENH). (B) Individual helices in ENH designs. The WT crystal structure is shown in CPK-based colors with carbon atoms in green. The design model for HT_ENH is shown with carbons in white. The design model for NSC is shown with carbons in pink. Electrostatic interactions, with less than 4 Å atom-atom distances, are shown with dotted yellow lines. The HT_ENH model has more intra-helical, short-range sidechain/sidechain than the other structures. (C) Sequences of ENH variants. Surface residues are in black text and other residues (which were not designed) are in gray. Among the surface residues, acidic sidechains in highlighted in light red, and basic sidechains are highlighted in light blue. The formal charge is given for each sequence.

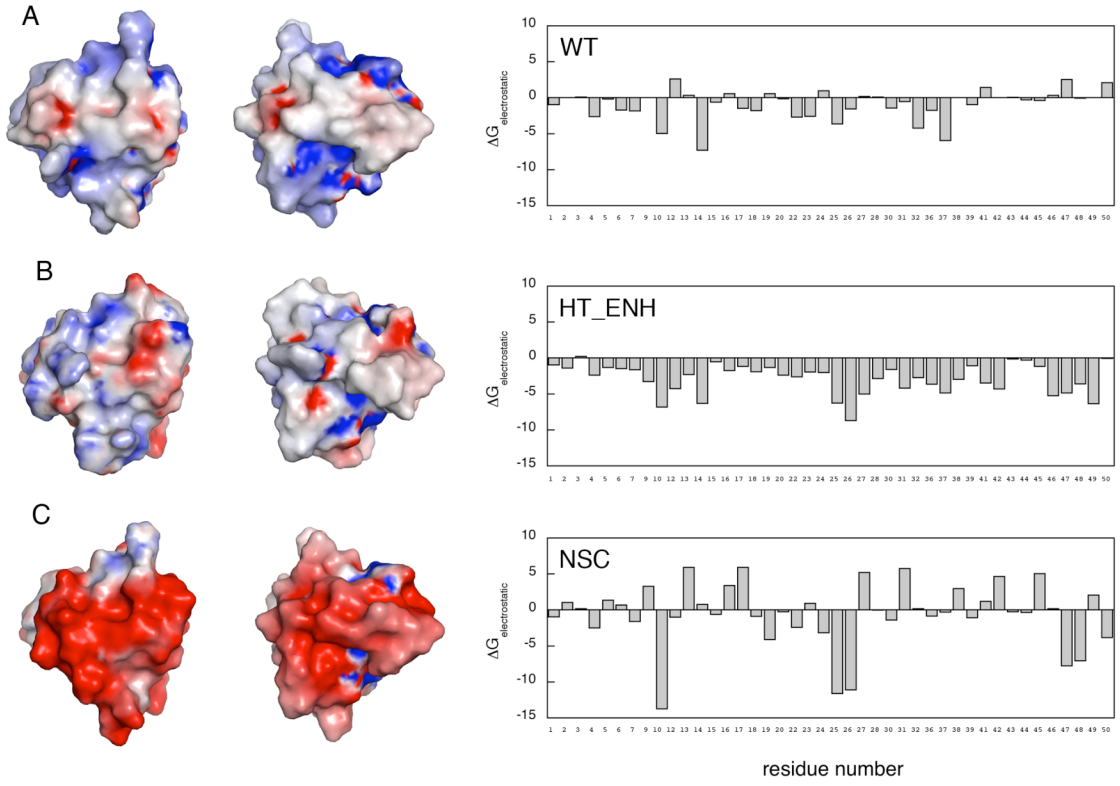


Figure 6-2. Electrostatic potential surface and calculated interaction energies for ENH variants: (A) WT, (B) HT_ENH, and (C) NSC. The coloring for the molecular surfaces is on a scale of ±5 kT. $\Delta G_{electrostatic}$ (in kcal mol^{-1}) is the sum of the solvent-screened Coulombic energy between the sidechain of interest and the backbone and also between the sidechain of interest and the other polar sidechains in the protein.

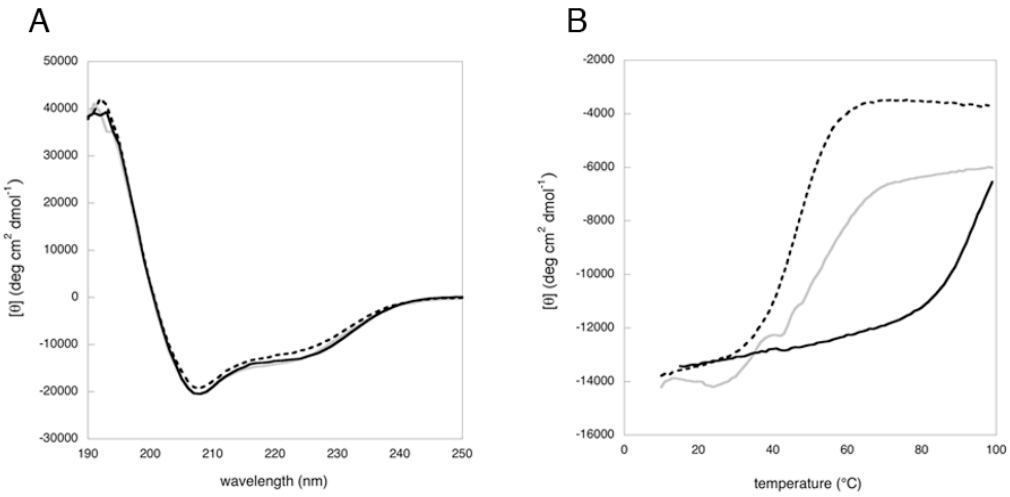


Figure 6-3. CD data for ENH variants. (A) CD wavelength scans. (B) Thermal denaturations. WT: gray; HT_ENH: black; NSC: dashed line. All data was collected at pH 6.5 in 50 mM sodium phosphate buffer. The thermal denaturation temperatures were calculated as the inflection points of the curves in (B): WT, 44°C; NSC, 48°C; HT_ENH, 94°C.

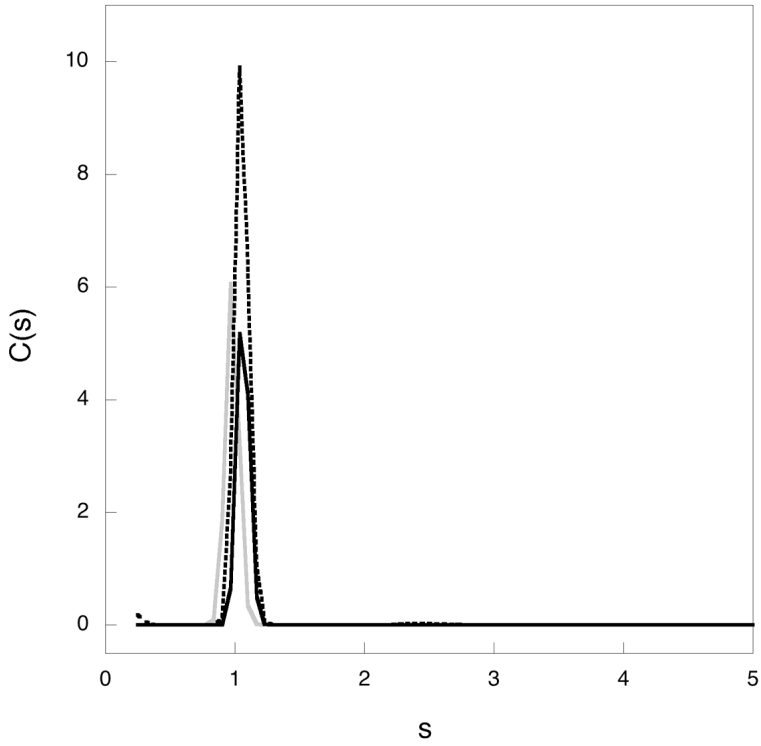


Figure 6-4. Analytical ultracentrifugation of HT_ENH. Gray line corresponds to WT ENH (75 μM). The solid black line is for HT_ENH at 75 μM protein concentration, and the dotted line is for HT_ENH at 150 μM protein concentration. Data was collected at pH 6.5 in 50 mM sodium phosphate and 150 mM NaCl.

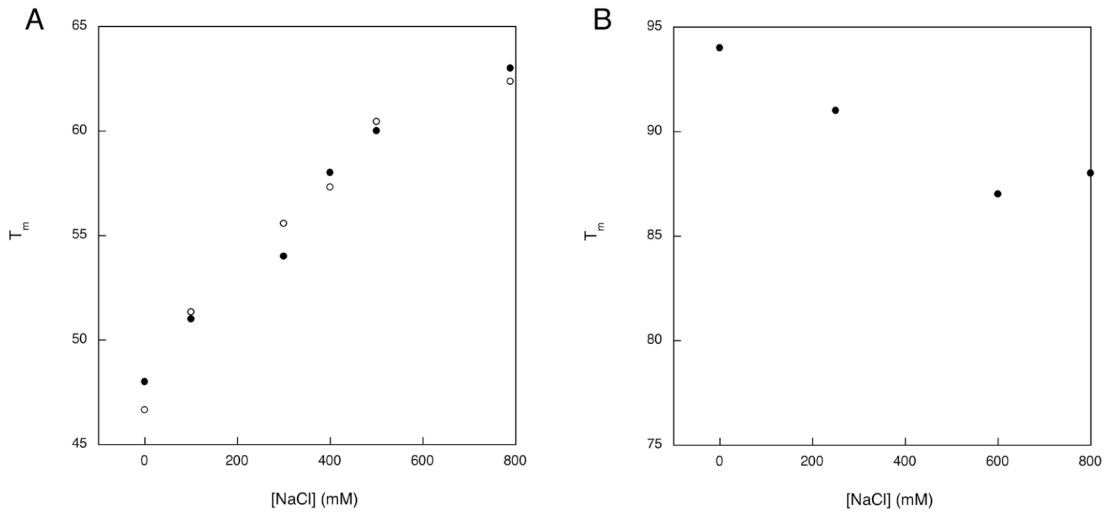


Figure 6-5. Effect of ionic strength on thermostability of (A) NSC and (B) HT_ENH. Open symbols correspond to T_m values calculated using non-linear fit to Equation 2 in Appendix A. The non-linear fit could not be calculated for HT_ENH due to a lack of post-transition baseline. Solid black circles correspond to the inflection point of thermal denaturation curves.

Chapter 7

Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function

Parts of this chapter were adapted from a published manuscript that was co-authored with Thomas P. Treynor, Daniel Nedelcu, and Professor Stephen L. Mayo

T.P. Treynor, C.L. Vizcarra, D. Nedelcu, and S.L. Mayo, *Proceedings of the National Academies of Science USA* **104**, 58–53 (2007).

**T.P.T. developed the DBIS algorithm. Experiments and analysis were a collaborative effort with T.P.T. and D.N.*

Abstract

In order to determine which of seven library design algorithms best introduces new protein function without destroying it altogether, seven combinatorial libraries of green fluorescent protein variants were designed and synthesized. Each was evaluated by distributions of emission intensity and color compiled from measurements made *in vivo*. Additional comparisons were made with a library constructed by error-prone PCR. Among the designed libraries, fluorescent function was preserved for the greatest fraction of samples in a library designed using a novel structure-based computational method. A trend was observed towards greater diversity of color in designed libraries that better preserved fluorescence. Contrary to trends observed among libraries constructed by error-prone PCR, preservation of function was observed to increase with a library's average mutation level among the four libraries designed with structure-based computational methods. Forty-one unique clones from a designed library were sequenced and consistent shifts in emission peak position were observed for three mutations in a variety of mutational backgrounds.

Introduction

Protein sequence space is so vast that one can easily imagine the optimal sequence for a particular application will never be sampled by random mutation and recombination. Structure-based computational protein design tools seek to screen that sequence space more thoroughly than can be screened in the laboratory, but are currently based on approximate representations of candidate sequences and an incomplete understanding of the relationships between structure and function. While many algorithms used to screen sequences *in silico* aim to identify a single optimal sequence,[1–5] others aim instead to optimize the composition of a library of sequences. [6–13] Provided resources exist to synthesize and screen such libraries, library design algorithms compensate for the approximations built into them by increasing the number of attempts at designing the desired function. Viewed from a complementary perspective, such algorithms aim to sample sequence space more effectively than methods that randomly generate sequence diversity.

Designed libraries can be synthesized for roughly the same cost as a designed sequence by recognizing the opportunities in gene synthesis for the combinatorial shuffling of sequence diversity.[14–17] Although many algorithms have now been proposed to design such combinatorial libraries,[7–9,11,12] few computationally designed libraries have been characterized experimentally,[9,18,19] and to our knowledge, there have been no controlled experiments comparing these methods with each other or with libraries of randomly generated sequence diversity. The results of such a comparison would be hard to predict, especially since none of these methods models protein function explicitly. Instead, these algorithms attempt to model protein stability as a surrogate for protein

function on the assumption that libraries with a greater fraction of well-folded proteins are more likely to contain variants with the desired function.

Here we evaluate seven designed combinatorial libraries of green fluorescent proteins (GFPs), including one with mutations picked at random. Preservation and diversity of function were judged using distributions of brightness and color, respectively, compiled from measurements made *in vivo* with a monochromator-based plate reader. GFP from *Aequorea victoria* modified by S65T [20] (GFP-S65T) was chosen as a reference sequence for each design algorithm because this variant is less extensively engineered than other variants whose structures have been solved to similarly high resolution. Positions 57 through 72 were targeted for this test because they form the longest contiguous stretch of core positions in the GFP-S65T structure.[21] The structure of GFP-S65T is illustrated in Figure 7-1A with the targeted positions shown in yellow. Because random core mutations are generally more disruptive than random surface mutations,[22,23] it was assumed that targeting core positions would provide better differentiation of designed libraries according to preservation and diversity of function criteria. Contiguity was imposed to allow the economical and high-fidelity cassette-based library synthesis. Where possible, libraries were controlled both for theoretical size and the precise distribution of mutation levels within each library since one would expect these factors to affect library quality when controlled for the same method of design.

We show that the corresponding design algorithms perform quite differently in this test. Four of the seven libraries were designed with structure-based computational methods: two with an algorithm introduced here (see Methods) and two with algorithms

described previously.[7,9] Among these four libraries, we observe that preservation of function increases with a library's average mutation level, contrary to the trends observed for libraries constructed by error-prone PCR (epPCR).[24,25] Across all seven libraries, we observe a trend towards greater diversity of function in designed libraries with greater preservation of function. An additional library generated by epPCR amplification of the entire GFP-S65T gene exhibited much less dispersion of function than designed libraries with similar preservation of function.

Results

Library Composition. The seven combinatorial libraries with compositions listed in Table 7-1 were designed, synthesized, and characterized as described in the Methods section. Briefly, the labels $DBIS^{ORBIT}$, $DBIS^{ORBIT}$ 4^4, C^{ORBIT}, and $SCMF^{ORBIT}$ 32^2 represent the four libraries designed using structure-based computational methods that draw on the ORBIT suite of protein design tools.[1–3] The $DBIS^{ORBIT}$ and $DBIS^{ORBIT}$ 4^4 libraries were designed using an algorithm whose principal innovations can be summarized as a diversity benefit applied to interacting sets of amino acids (DBIS). The C^{ORBIT} library was designed with a consensus method (C) based on the work of Hayes *et al*.[9] The $SCMF^{ORBIT}$ 32^2 library was designed using a self-consistent mean field (SCMF) calculation to direct combinatorial saturation mutagenesis as suggested by Voigt *et al*.[7] The C^{MSA} and SE/C^{MSA} libraries were each designed with the same multiple sequence alignment (MSA) of naturally occurring fluorescent proteins.[26] Both use a consensus method derived from the work of Hayes *et al*. [9], but the latter is distinguished by directing mutations to positions that have the largest site entropies (SE). Mutations in the Random library were picked with a

random number generator. In order to approach 95% confidence that the true extremes of function in each library would be sampled, we aimed to sample most designed libraries by three times their theoretical size.[27] Considering also that one-half hour was needed to acquire each set of 96 high-resolution emission spectra, these constraints dictated that theoretical library sizes should be close to 500. Although this size is orders of magnitude smaller than most libraries screened for binding[28] or low-resolution fluorescence properties,[29,30] it is especially relevant to difficult-to-screen functions such as improved enzymatic activity with non-fluorogenic substrates. It was assumed that the best differentiation between design algorithms would be achieved by applying them in ways that maximized the average number of mutations per sequence, yet each combinatorial library was constrained to include the sequence of GFP-S65T so that none would be rendered completely non-functional due to a uniquely disruptive mutation. Thus most designed libraries tested here ($DBIS^{ORBIT}$, C^{ORBIT}, C^{MSA}, SE/C^{MSA}, and Random) have a theoretical size of 2^9 and an average of 4.5 mutations per sequence. The $DBIS^{ORBIT}$ 4^4 and $SCMF^{ORBIT}$ 32^2 libraries have unique sizes and average mutation levels that are conveyed by the labels we have given them. For example, the $SCMF^{ORBIT}$ 32^2 label indicates that this library was made by combinatorial saturation mutagenesis at two positions using 32-fold degenerate codons.

It is interesting to note the extent to which the compositions of the designed libraries reflect the fact that evolution disfavors ionizable side chains in protein cores. The MSA used to design the C^{MSA} and SE/C^{MSA} libraries illustrates this trend, with a notable exception being the unusually high degree of conservation at position 69 for a buried basic side chain.[26] The scoring function used for structure-based design was

parameterized specifically to prevent the desolvation of hydrophilic side chains in protein cores under most circumstances.[31] Thus the DBIS^{ORBIT} 4^4 library introduces only one acidic side chain among its twelve mutations distributed over four positions, and the DBIS^{ORBIT} and C^{ORBIT} libraries do not introduce any ionizable side chains anywhere. Although the SCMF^{ORBIT} 32^2 library was designed using the same scoring function as these three other libraries, imposing saturation mutagenesis for this one library makes it introduce many mutations that are strongly disfavored by this scoring function. Thus the SCMF^{ORBIT} 32^2 library introduces ionizable side chains at core positions with greater frequency than each library tested except the Random library.

Preservation of Function. For each of the designed libraries, and for the epPCR library, emission spectra were recorded for roughly 1500 bacterial cultures expressing GFP variants. We define the brightness and color of each spectrum sampled as its integrated emission intensity and average position, respectively. Because it is not clear how best to define a functional sample, we have quantified each library's preservation of function in three ways. For each library, the percentage of samples that have at least one-half, one-tenth, and one-fiftieth the brightness of cultures expressing GFP-S65T are presented as bar graphs in Figure 7-2. By all three of these measures, most of the designed libraries performed considerably better than the Random library. Only 1.6% of samples from the Random library had at least one-fiftieth the brightness of cultures expressing GFP-S65T. Although the SCMF^{ORBIT} 32^2 library had a larger fraction of functional samples than the Random library by this most inclusive definition of function, it had a similar fraction by the most exclusive definition. The relatively poor performance of these two libraries is

probably due in part to the relatively large frequencies with which these libraries introduce ionizable side chains to the protein core.

By all three of these measures the DBIS^{ORBIT} library performed best of all. More than 10% and 40% of its samples were at least one-half and one-fiftieth as bright as cultures expressing GFP-S65T, respectively. The C^{ORBIT} library performed nearly as well. The SE/C^{MSA}, C^{MSA}, and DBIS^{ORBIT} 4^4 libraries performed similarly to each other, with close to 1% and 10% of samples being at least one-half and one-fiftieth as bright, respectively, as cultures expressing GFP-S65T. The Q69R mutation, since it introduces an ionizable side chain to the protein core, would seem responsible for much of the weaker performance of the MSA-based libraries compared to the DBIS^{ORBIT} and C^{ORBIT} libraries, which instead introduce the Q69L mutation. However, even if it is assumed that the Q69R mutation always disrupts function and that the Q69L mutation never disrupts function, less notable differences among these libraries must account for at least half the observed differences in performance.

Multiple epPCR libraries were synthesized using different mutation rates. Only the library that appeared to have a similar fraction of functional samples as the DBIS^{ORBIT} library was characterized in detail in order to compare average mutation levels and diversity of function under this condition. Despite the fact that random mutations are generally tolerated at surface positions better than at core positions,[22,23] the average number of non-synonymous mutations for genes in this epPCR library was determined by DNA sequencing to be 2.5, roughly half the average of 4.5 mutations per gene for the core-directed DBIS^{ORBIT} library.

Diversity of Function. Because the dimmest samples have colors biased by emission from molecules other than GFP, here we consider only those samples with at least one-half the brightness of cultures expressing GFP-S65T. Of the 11,575 spectra sampled, 701 met this criterion. The red-most and blue-most of these spectra are illustrated in Figure 7-1B.

The diversity of function for a library of fluorescent proteins may be associated with either its extremes of color or its dispersion of color. The former we define as the difference between the positions of the red-most and blue-most spectra in a library. Figure 7-3 illustrates the set of colors sampled for each library with black marks, such that the separation between left-most and right-most marks illustrates a library's performance according to this extremes-of-function metric. We define dispersion of function as the difference between the positions of the spectra that lie one quartile above and below the median for a library. In Figure 7-3, this median is illustrated with a white bar on top of a red box illustrating the positions of the first and third quartiles.

The seven designed libraries are thus seen to cluster into four performance categories based on these complementary metrics. The $DBIS^{ORBIT}$ and C^{ORBIT} libraries outperform all the other designed libraries by having both the largest separation between extremes and the greatest dispersion. The SE/C^{MSA} and C^{MSA} libraries constitute the next category by having greater separation between extremes than the $DBIS^{ORBIT}$ 4^4 and Random libraries, though similar dispersion. The $SCMF^{ORBIT}$ 32^2 library then constitutes the last category by having both the smallest separation between extremes and the least dispersion. By the extremes-of-function metric, the epPCR library performs better than each of the designed libraries except the $DBIS^{ORBIT}$ and C^{ORBIT} libraries; however, by the

dispersion-of-function metric, the epPCR library performs worse than each of the designed libraries except the $SCMF^{ORBIT}$ 32^2 library.

A complementary illustration of the preservation and diversity of function sampled from each library is provided in Figure 7-4. For each library, the width of each spectrum sampled is plotted against its color with a circle of area proportional to its brightness. Although Figure 7-4 does not characterize the libraries with the statistical rigor of Figures 7-2 and 7-3, it does provide additional support for the clustering and ranking of the designed libraries described above. It also reveals a striking correlation between emission line shape and emission color among the brightest samples in each library. We have investigated the physical mechanisms that may be responsible for this trend with additional measurements that will be presented elsewhere Treynor *et al*. (in preparation).

Mutational analysis. The 96 brightest samples from the $DBIS^{ORBIT}$ library were sequenced, providing 41 unique mutants of GFP-S65T. In addition, several mutants were constructed to complete quadruple mutant cycles to be reported elsewhere Treynor *et al*. (in preparation). This "synthetic sequence family" for GFP is shown in Figure 7-5A. For each mutation in the $DBIS^{ORBIT}$ library, the average affect of mutating that residue in the background of many different sequences was evaluated. In Table 7-2, the average shifts in peak position, peak width, Stoke's shift and apparent T_m are reported for each mutation. There are three mutations that show robust shifts in peak positions in many mutational backgrounds: T65A, Q69L, and S72A. The first two mutations cause blue-shifted and broadened emission spectra whereas the latter causes red-shifted and

narrowed emission spectra. These residues are highlighted in Figure 7-5B. The sidechain of Ser72 is far from the chromophore, and residue 65 is part of the chromophore backbone, indicating that mutations at these residues might act through the backbone of the helix to alter fluorescence properties of the chromophore. The sidechain of Gln69 is pointed away from the chromophore but makes contact with several ordered waters near the chromophore (Figure 7-5B). The mutations that caused significant change in Stoke's shift were T62A, V68A, and S72A, indicating that the blue-shifting mutations at residues 65 and 69 are not operating by reducing the Stoke's shift. The largest effects on stability (as judged by apparent T_m) were observed for Q69L, which is stabilizing, and T62A, which is destabilizing. Most mutations had negligible average effects on apparent stability ($\Delta T_m \sim 1–2$ °C).

Discussion

Figure 7-2 illustrates that preservation of function increases with average mutation level among the four libraries designed using structure-based computational methods. The opposite trend has been observed for protein libraries synthesized by epPCR,[24,25] and would suggest that, constrained to a particular library size, the designed library with the lowest mutation rate should yield the largest fraction of functional samples. It is thus notable that the poor performance of the $SCMF^{ORBIT}$ 32^2 library in this respect may have more to do with the overarching strategy that enforced its low mutation rate, combinatorial saturation mutagenesis, than the computational method used to select positions for mutation. A library defined by combinatorial saturation mutagenesis would

have to tolerate roughly 12 different amino acids per position to preserve function as well as the DBIS^{ORBIT} and C^{ORBIT} libraries. Finding any two core positions in GFP-S65T that could accept such great diversity, let alone two between positions 57 and 72, would seem an especially difficult problem.

Figure 7-3 illustrates that diversity of function tends to increase with preservation of function among the seven designed libraries. This result justifies an approach to library design where protein stability is modeled as a surrogate for protein function,[7–9,11,12] as long as mutations are directed towards positions likely to perturb function. Moreover, this result suggests that improvements in modeling protein stability should yield designed libraries that sample a wider array of protein functions.

A frequently desired trait among GFP variants has been red-shifted emission.[29,32,33] Although the vast majority of the bright variants sampled from the epPCR library have emission spectra nearly identical to cultures expressing GFP-S65T, the one sample from this library with a substantial red-shift did have the red-most spectrum sampled in our test. The corresponding GFP gene was sequenced and determined to have the V224I and M233K mutations. Only the V224I mutation is in the core of the protein and close to the chromophore, suggesting that it is primarily responsible for the observed red-shift. The fact that neither of these mutations involves the positions targeted in the test underscores the way the performance of a designed library is intrinsically limited by the quality of the information in the design, such as the choice of positions targeted for mutation. Nevertheless, the far greater number of almost identically red-shifted samples from the DBIS^{ORBIT} and C^{ORBIT} libraries indicates that our best information at present is a valuable tool with which to complement epPCR for sampling diverse functions.

Even though red-shifted emission is frequently desired for GFPs, other measures described here may be more relevant to the extrapolation of these results to other protein engineering projects. Such projects typically aim to increase the stability of an enzyme, its rate of catalysis, or the affinity of a protein for a ligand.[28,34] Since denatured GFP does not fluoresce,[35] one interpretation of Figure 7-2 is that the algorithms that preserved function best did so by disrupting the global structure of GFP the least. According to this interpretation, we would predict that the algorithms used to design the $DBIS^{ORBIT}$ and C^{ORBIT} libraries would also perform best when attempting to stabilize an enzyme with core-directed mutations. However, the relative performance of the MSA-based methods might be expected to increase in this case if the covariances among amino acid frequencies important for protein stability can be extracted from evolutionary noise.[13,36,37]

The emission spectrum of GFP is a reporter on the local structure of its chromophore. In other words, a more varied sampling of spectral properties is equivalent to a more varied sampling of structures at the "active site" of GFP. Thus, based on Figures 7-2 and 7-3, we can predict that the algorithms used to design the $DBIS^{ORBIT}$ and C^{ORBIT} libraries will provide the most diverse sampling of active site structures in functional enzymes. Structure-based computational methods should thus prove especially useful for relatively low-throughput screening projects in which libraries made by epPCR, even those with low mutation rates, cannot be screened thoroughly.

In summary, we have shown that small combinatorial libraries can exhibit considerable diversity of function if designed well. Based on the design and results of this test, we recommend complementing more widely used strategies for generating functional diversity such as epPCR and combinatorial saturation mutagenesis with a

strategy that defines a combinatorial library by a single conservative mutation at each of many positions close to a protein's active site. We have found structural information as utilized by the DBIS algorithm or the method of Hayes *et al*.[9] to be more successful than limited evolutionary information in identifying compatible conservative mutations. Although currently limited by the need for an accurate structure, the utility of the structure-based design algorithms should improve as methods improve for docking ligands onto proteins and for determining protein structures from protein sequences. Indeed the great promise of these methods for library design is that they might be used to implement a knowledge-based approach to engineering totally novel functions for which no natural protein exhibits even the slightest glimmer of the desired function. In the meantime, the mutational analysis presented here shows that this approach to protein engineering should prove especially useful for investigations of protein structure-function relationships, where ideally large numbers of differently functional variants would be related by the same small set of mutations.

Methods

Rotamer energies. Rotamer energy calculations were based on a 1.45 Å-resolution structure of *A. victoria* GFP containing the S65T and Q80R mutations (PDB code 1q4a).[21] The united residue "CRO" at position 66 was broken up into three residues (positions 65–67), and atoms were renamed according to standard conventions. Hydrogen atoms were added to the protein and side chains flipped as suggested by MOLPROBITY.[38] Hydrogen atoms on the chromophore were hand-edited using BIOGRAF (Molecular Simulations, Inc.). Waters in the structure were removed.

The resulting structure underwent conjugate gradient minimization for 50 steps using the DREIDING force field,[39] without electrostatics. In order to conduct the electrostatics calculations described below, partial charges for residue 66 were assigned to values derived by Helms et al. from Restricted Hartree Fock calculations of the S_0 state of the anionic chromophore.[40] All other partial charges were as defined by the PARSE parameter set,[41] except the N-terminal nitrogen was assigned a partial charge of 0, and the carboxyl carbon of T65 was assigned a partial charge of 0.67 to achieve a total charge of –1.00 for residues 65–67.

Rotamer singles and pairs energies were calculated using a scoring function with terms for van der Waals interactions (E_{vdw}), hydrogen bonding ($E_{h\text{-}bond}$), electrostatics (E_{elec}), and atomic solvation (E_{as}):

$$E_{total} = E_{vdw} + E_{h-bond} + E_{elec} + E_{as}$$

Van der Waals energies were calculated according to

$$E_{vdw} = D_0\left[\left(\frac{\alpha R_0}{R}\right)^{12} - 2\left(\frac{\alpha R_0}{R}\right)^{6}\right]$$

where R is the interatomic distance between two atoms, D_0 is the geometric mean of the well depths of the two atoms, R_0 is the geometric mean of the van der Waals radii of the two atoms, and α is a van der Waals radius scaling factor[42] set equal to 0.9. Hydrogen bonding energies were calculated according to

$$E_{h-bond} = D_0\left[5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10}\right]F\left(\theta,\phi,\varphi\right)$$

where R is the distance between hydrogen donor and acceptor atoms, D_0 is a hydrogen bond well depth set equal to 8 kcal/mol, R_0 is a hydrogen bond equilibrium distance set

equal to 2.8 Å, and $F(\theta,\phi,\varphi)$ is a geometric factor defined elsewhere.[2] The hydrogen bonding energy for any hydrogen bond between a side chain rotamer and the backbone atoms in the same residue was set to zero. Electrostatics energies were calculated according to

$$E_{elec} = \frac{qq'}{\varepsilon R}$$

where q and q' are the partial charges of two atoms, R is the distance between them and ε is a dielectric constant set equal to $40R$. Atomic solvation energies were calculated according to

$$E_{as} = -(\kappa + 1)\sigma_{np} A_{np,b} + \kappa\sigma_{np} A_{np,e} + \sigma_p A_{p,b}$$

where $A_{np,e}$ is nonpolar exposed surface area, $A_{np,b}$ is nonpolar buried surface area, $A_{p,b}$ is polar buried surface area, κ is a nonpolar exposure scale factor set equal to 1.6, σ_p is a scale factor set equal to 0.1 kcal/mol/Å^2 that penalizes polar burial, and σ_{np} is a scale factor set equal to 0.026 kcal/mol/Å^2 that benefits and penalizes nonpolar burial and exposure, respectively. Solvent-accessible surface areas were calculated using the Connolly algorithm as described elsewhere.[31,43]

The May 2002 version of Dunbrack's backbone-dependent rotamer library[44] was expanded by rotation of ±1 standard deviation about χ_1 and χ_2 for every rotamer. Rotamer singles energies, $E_{rot}(i_r)$, were evaluated for each of these rotamers r at each position i in the set [57–65, 67–72], except for cysteine and proline rotamers. At each position i the singles energy for the rotamer defined by the conjugate-gradient minimized structure, $i_{current}$, was also evaluated. Any rotamer with a singles energy greater than 20 kcal/mol was eliminated from the rest of the calculation. Rotamer pairs energies,

$E_{rot}(i_r,j_s)$, were then calculated for the remaining rotamers r and s at positions i and j, respectively.

The DBIS Algorithm. One of the fundamental innovations of the DBIS algorithm is that it aims to explicitly model the interactions among sets of amino acids at the positions targeted for design. Set singles and pairs energies are constructed analogous to rotamer singles and pairs energies in structure-based computational protein design.[1] Thus the exact optimization algorithms used to determine the global minimum energy conformation (GMEC) from a rotameric representation of the sequence design problem[45,46] can be used instead to determine the global minimum energy combinatorial library (GMEL) from a set-based representation of the combinatorial library design problem.

Figure 7-6 illustrates the main components of the generalized DBIS algorithm. A symmetric matrix of rotamer singles and pairs energies is first calculated using a template structure and rotamer library.[1–3] This rotameric representation of the sequence design problem is then projected onto a smaller matrix with one row and one column for each combination of amino acid and targeted position (*vide infra*). These amino acid singles and pairs energies are then combined to build the set-based representation of the combinatorial library design problem by filling a matrix with one row and one column for each set of amino acids considered at each position in the library design.

Here we have implemented the generalized DBIS algorithm such that a library's energy is equal to an arithmetic average of conformational energies calculated for each sequence in the library, adjusted for composition and diversity benefits. Optimizing

library composition thus corresponds to minimizing this energy. For rotamer r at each position i, the energy of point mutation, $E_{pm}(i_r)$, is evaluated as

$$E_{pm}(i_r) = E_{rot}(i_r) + \sum_{j \neq i} E_{rot}(i_r, j_{current})$$

where $E_{rot}(i_r)$ and $E_{rot}(i_r, j_{current})$ are rotamer singles and pairs energies, respectively, and $j_{current}$ is the rotamer defined by the amino acid at position j in the template structure. Within the set of rotamers r at position i corresponding to amino acid a, $i_r \in i_a$, the rotamer that minimizes $E_{pm}(i_r)$ is represented as $i_{min,a}$. If there exists some $i_r \in i_a$ that has survived the previous rotamer pruning step (*vide supra*), the amino acid singles energy for amino acid a at position i, $E_{aa}(i_a)$, is then set equal to

$$E_{aa}(i_a) = E_{rot}(i_{min,a}) + E_{comp}(i_a)$$

where the composition benefit $E_{comp}(i_a)$ has a user-defined value that biases optimization towards or away from libraries that include amino acid a at position i. Otherwise $E_{aa}(i_a)$ is set equal to the cutoff value used to prune rotamers, 20 kcal/mol, such that these amino acids are effectively eliminated from the calculation; a value similar to some of the better rotamer singles energies could conceivably improve library design for some applications by complementing the conservative nature of our structure-based method with a desired degree of randomness. Assignment of the amino acid energies in this manner effectively prunes the rotamers in the calculation to no more than one rotamer per amino acid per position.

If there exists some $i_r \in i_a$ and some $j_s \in j_b$ that have survived the rotamer pruning step, the amino acid pairs energy, $E_{aa}(i_a,j_b)$, is then set equal to

$$E_{aa}(i_a, j_b) = E_{rot}(i_{min,a}, j_{min,b}) \ .$$

Otherwise $E_{aa}(i_a,j_b)$ is set equal to the cutoff value used to prune rotamers, 20 kcal/mol, such that these amino acids are effectively eliminated from the calculation; a value similar to some of the better rotamer pairs energies could conceivably improve library design for some applications by complementing the conservative nature of our structure-based method with a desired degree of randomness.

For the set of amino acids a represented by x, a set singles energy, $E_{set}(i_x)$, is calculated at each position i as

$$E_{set}(i_x) = \frac{1}{N_x}\left[\sum_{a\in x} E_{aa}(i_a)\right] - L\cdot\ln(N_x)$$

where N_x is the number of amino acids in set x, and L is a factor used to control the size of the optimal library. We refer to the second term in this equation as a diversity benefit and to L as a diversity benefit scale factor. Faced with two libraries of the same size, the logarithmic form of the diversity benefit will tend to favor the one with sequence diversity distributed over a greater number of positions. A quadratic form would have the opposite effect and may be more desirable depending on one's application. Of course, the functional form for the diversity benefit is inconsequential when only two set sizes are considered in a design, as was the case in designing the DBIS^{ORBIT} and DBIS^{ORBIT} 4^4 libraries (see below). For sets x and y at positions i and j, the set pairs energy is then calculated as

$$E_{set}(i_x,j_y) = \frac{1}{N_xN_y}\left[\sum_{a\in x}\sum_{b\in y} E_{aa}(i_a,j_b)\right] \quad .$$

The composition of the optimal combinatorial library was thus defined by the optimal combination of these set singles and pairs energies. In designing the DBIS^{ORBIT} and DBIS^{ORBIT} 4^4 libraries, we first imposed $E_{comp}(i_a) = 0$ at all positions; if the GMEL for the

value of L that gives the desired library size did not include the GFP-S65T sequence, we iteratively altered $E_{comp}(i_a)$ in –5 kcal/mol increments for the missing GFP-S65T residues until this sequence was recovered in the designed library.

Library Design Methods. Composition, set size and genetic code constraints were enforced for all tested design algorithms to facilitate comparisons among them. The genetic code constraint allowed each library to be constructed at minimal cost and effectively applied some of the physicochemical information that may exist in the genetic code to the process of design (it is notable that there were large differences in performance among libraries although each shared this constraint). Relaxing the genetic code constraint would change the composition of each designed library substantially and could alter the observed performance ranking.

One set of rotamer singles and pairs energies was used in four different ways to design the DBIS^{ORBIT}, DBIS^{ORBIT} 4^4, C^{ORBIT}, and SCMF^{ORBIT} 32^2 libraries. In order for the DBIS algorithm to yield a library of 2^9 sequences that included GFP-S65T, all values of $E_{comp}(i_a)$ were set equal to zero except $E_{comp}(63_T)$ = –10 kcal/mol and $E_{comp}(69_Q)$ = –5 kcal/mol; the only sets considered at each position were the 95 unique sets of either one or two amino acids that can be defined by the use of mixed bases during primer synthesis; L was set equal to 6.5. In order for the DBIS algorithm to yield a library of 4^4 sequences that included GFP-S65T, all values of $E_{comp}(i_a)$ were set equal to zero except $E_{comp}(63_T)$ = –10 kcal/mol and $E_{comp}(69_Q)$ = –10 kcal/mol; the only sets considered at each position were the 113 unique sets of either one or four amino acids that can be defined by the use of mixed bases during primer synthesis; L was set equal to 4.6.

The SCMF^{ORBIT} 32^2 library was designed by applying the method of Voigt *et al.* [7] in the following way. Each rotamer was first assigned a probability equal to the inverse of the number of rotamers at its position. The self-consistent mean-field solution was then calculated for an initial temperature of 50,000 K. As the temperature was lowered in 100 K increments, the solution from each previous temperature was used as the initial configuration for the next temperature. Saturation mutagenesis was directed to the two positions with site entropies greater than 1.0 at a final temperature of 1000 K.

The C^{ORBIT} library was designed by applying the consensus method (C) of Hayes *et al.* [9] in the following way. The GMEC for this design problem was used as the initial configuration for a Monte Carlo trajectory through conformation space. One million steps were used for each of 100 cycles during which temperature oscillated between 4000 K and 150 K. Only the 1010 unique amino acid sequences with the best energies sampled were retained for further analysis. At 9 of 15 positions, there appeared at least one mutation that could be introduced to GFP-S65T by a single nucleotide substitution. The C^{ORBIT} library was thus defined by the one such mutation that appeared with the greatest frequency at each of these nine positions. (At 1000 sequences a unique library could not be defined by this method since both alanine and threonine appeared with equal frequency at position 58.) Three apparent deficiencies of this consensus method were addressed by developing the DBIS algorithm: first, Monte-Carlo-based sampling of the energy landscape is by its nature both inexhaustive and random; second, disruptive combinations of amino acids might arise when a library is designed without accounting for correlations in an alignment; and third, even if correlations were accounted for, any

alignment with enough sequences to truly reflect global trends in these correlations would likely be too large to be practical.

The C^{MSA} and SE/C^{MSA} libraries were each designed with the same alignment of naturally occurring fluorescent proteins according to similar consensus methods. Of the 48 GFP homologs aligned by Shagin *et al.*,[26] we used only the 36 homologs labeled as either GFPs, YFPs, CyFPs, or RFPs. To design the C^{MSA} library, a consensus method (C) derived from the one employed by Hayes *et al.* [9] was used. At 12 of the positions between 57 and 72 there appeared at least one mutation that could be introduced to GFP-S65T by a single nucleotide substitution. The nine positions that had at least one such mutation represented at least four times were mutated to whichever of these mutations occurred with the greatest frequency at each position. Because two such mutations occurred with greatest frequency at positions 62 and 72, we elected in each case to introduce the mutation that happened to be shared with the $DBIS^{ORBIT}$ library. The approach used to design the C^{MSA} library thus directs mutations away from the positions that exhibit the least conservation. To explore the possibility that these least conserved positions might tolerate mutation best, the SE/C^{MSA} library was designed by directing mutations to the 9 positions (of 12) that had the greatest site entropies,

$$s_i = -\sum p(i_a) \ln p(i_a)$$

where $p(i_a)$ is the frequency of amino acid a at position i, and the sum is taken over all amino acids for which $p(i_a) \neq 0$. The mutations introduced at these positions were chosen by the same considerations used to design the C^{MSA} library. We did not use any design algorithms that utilized pairwise correlations among the mutations in the MSA since this

alignment was rather small and there may be considerable evolutionary noise in such correlations.[36,37]

The Random library was designed using a Python script to pick one mutation at random at each of the nine positions mutated in the DBIS^{ORBIT} library.

Library Synthesis and Characterization. Designed libraries were synthesized with a cassette-based method derived from that of Hiraga and Arnold for site-directed recombination.[17] The gene for GFP-S65T was first constructed between unique *Sfi*I recognition sequences by gene assembly[47] and inserted into a vector derived from pBAD18-Cm.[48] The sequence for positions 57–72 was then replaced with a restriction fragment and recognition sequence for the Type IIB restriction enzyme *Bsa*XI using site-directed mutagenesis by overlap extension,[49] followed by digestion of the modified gene with *Sfi*I (New England Biolabs #R0123S) and ligation into the vector with T4 DNA ligase (New England Biolabs #M0202S). The resulting vector is mapped in Figure 7-7. The non-coding strand of the *Bsa*XI restriction fragment was designed to have both a stop codon and a *Blp*I recognition sequence. The modified vector was digested with *Bsa*XI (New England Biolabs #R0609S), purified by spin column and dephosphorylated with CIP (New England Biolabs #M0290S). The large restriction fragment was then purified by gel extraction.

Degenerate primers corresponding to the library designs in Table 7-1 were ordered from Integrated DNA Technologies. Each degenerate codon was selected to minimize degeneracy and maximize codon usage in *E. coli*. Complementary degenerate primers were dissolved in 10 mM Tris buffer pH 8.5 to a concentration of 100 μM. 1 μL

of each solution was combined and diluted to a final volume of 20 μL in T4 polynucleotide kinase (PNK) buffer (New England Biolabs #M0201S). Primers were annealed by heating for 2 min at 95 °C and cooling to room temperature on the bench. 1 μL 0.1 M DTT, 1 μL 0.1 M ATP, and 10 U PNK were added to solution. This phosphorylation reaction was conducted for 3 h at 37 °C. PNK was deactivated by incubation at 65 °C for 20 min.

A ligation reaction was conducted by mixing 1 μL of a 10-fold dilution of the phosphorylation reaction with roughly 100 ng of the gel-extracted vector fragment in 20 μL T4 ligase buffer with 400 U T4 DNA ligase. This ligation reaction was incubated at room temperature for 1 h. The ligase was removed and DNA eluted into 30 μL 10 mM Tris buffer pH 8.5 using QIAGEN's QIAquick PCR Purification Kit. 28 μL of this solution was then reacted with 20 U *Blp*I (New England Biolabs #R0585S) for 1 h at 37 °C to cut open any vector still containing the small *Bsa*XI restriction fragment. *Blp*I was removed and DNA eluted into 30 μL water using QIAGEN's QIAquick PCR Purification Kit.

The epPCR library was constructed in a 100 μL reaction mixture containing 20 ng vector containing the gene for GFP-S65T, 0.5 μM forward primer (5'-CCTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTC – 3'), 0.5 μM reverse primer (5'- TCTTCTCTCATCCGCCAAAACAGCCAAGCTTGCATGCCTG – 3'), 7 mM $MgCl_2$, 400 μM $MnCl_2$, 500 μM dTTP and dCTP, 200 μM dATP and dGTP, 1x Applied Biosystems PCR Buffer II without $MgCl_2$, and 5 U Applied Biosystems AmpliTaq DNA Polymerase. The PCR program consisted of 95 ºC for 5 min followed by 14 cycles of 30 s each at 95 ºC, 50 ºC, and 72 ºC. The PCR product was purified and

eluted with 40 μL 10 mM Tris buffer pH 8.5 using QIAGEN's QIAquick PCR Purification kit. DNA was then digested with 20 U *Sfi*I (New England Biolabs #R0123S) in a 50 μL volume. The heaviest restriction fragment (~ 1 kb) was cut out and purified using QIAGEN's QIAquick Gel Extraction Kit. 20 ng of the purified insert was mixed with 80 ng of *Sfi*I-digested vector and 400 U T4 DNA ligase in a 20 μL volume for ligation at room temperature for 1 h.

Electrocompetent cells were prepared from *E. coli* strain NM554 purchased from Stratagene. Ligation reactions were transformed with these cells and spread onto agar plates containing LB medium, 34 μg/mL chloramphenicol (Cm, Sigma #C0378) and 0.2% L-(+)-arabinose (Ara, Sigma #A3256). Seven colonies from each designed library were picked for sequencing to judge whether or not each library sufficiently resembled its design. All mutations observed were as designed, and each of the mutations in the libraries of 2^9 sequences were observed at least once. In order to quantify the mutation rate in the epPCR library, nine genes were sequenced. Each sequence contained at least one mutation, and no insertions or deletions were observed. A rate of 3.4 ± 0.6 mutations per gene (standard error calculated assuming Poisson counting statistics) was estimated from the 31 mutations (23 non-synonymous) observed among 6453 bases.

For each library, individual colonies were picked on the bench top with sterile toothpicks into 18 sterile 96-well plates (Nunc #263339) containing 250 μL LB/Cm. Colonies of every morphology were picked except the few colonies that were distinctly larger and whiter than the rest. Roughly 3% of all colonies had this morphology, even when a "library" of only GFP-S65T was prepared identically. The following controls were built in to each plate: wells A6-D6 were inoculated with colonies of NM554

bacteria expressing GFP-S65T; wells E6-H6 were inoculated with colonies of NM554 bacteria expressing GFP truncated at position 56; wells B2, B11, G2 and G11 were not inoculated.

Plates were covered with breathable sealing tape (Nunc #249720) and a plastic lid (Nunc #249944). Cultures were grown for 24 h by shaking at 250 rpm at 30 °C (Labline #3525). A 96-pin replicator (Nunc #250520) was then used to inoculate 250 μL LB/Cm/Ara with the starter cultures. After shaking 24 h, these expression cultures were pelleted by centrifugation at 5100 rpm using a tabletop centrifuge (Beckman-Coulter S5700 Rotor and Allegra 25R Centrifuge). Supernatant was decanted and pellets were washed with two cycles of resuspension in 300 μL PBS buffer, centrifugation and decanting. Pellets were refrigerated for 5 days at 4 °C to allow some of the variants with slower rates of chromophore maturation to mature.

Pellets were then resuspended in 250 μL PBS buffer, and 200 μL was transferred to black plates with a clear bottom (Greiner #655096) for absorption and emission measurements using a monochromator-based plate reader (Tecan Safire). First, optical density was recorded at 600 nm (OD_{600}) to gauge variations in sample handling across wells in the same plate. Samples prepared as described typically had absorbance near 0.5. Second, the optimal detector gain for a plate was determined by exciting samples at 460 nm with a bandwidth of 12 nm and measuring emission at 510 nm with a bandwidth of 2.5 nm. This gain (with a value typically between 80 and 85 units) was then used as emission spectra were recorded from 475 to 599 nm with 2 nm steps.

To prepare proteins expressed in 96-well format for melting measurements using a QPCR instrument, cultures were first washed twice with PBS buffer and frozen at –

80°C. Cells were then thawed and resuspended in lysis buffer (100 mM NaPi pH 8.0, 10 mM MgCl2,1 U/mL DNase, 0.5 mg/mL lsyozyme). After 1 h at 37°C plates were centrifuged for 15 min at 5700 rpm. Lysates were separated from cell debris and diluted individually with PBS buffer to allow melt curves to be recorded with similar precision for variants initially present at different concentrations. 50 μL aliquots of each dilution were used for melting measurements. Just prior to these measurements well factors were determined using a separate plate with 50 nM fluorescein in each well in order to avoid the unfolding of protein that would have accompanied an internal well factor calibration. Melt curves were then compiled from 30 to 100°C in increments of 1°C. Samples were given 2.5 min at each temperature before the plate was imaged briefly with a CCD camera. The apparent T_m value was taken to be the inflection point of the melting curve.

Analysis. MATLAB scripts were written to calculate each spectrum's integrated intensity, average position, and line width in the following manner. Spectra were first truncated to values measured between 481 and 599 nm in order to avoid artifacts from leakage of the light source into the detector at the shortest wavelengths. In order to report positions and widths that more closely reflect the intrinsic energy spectra of the molecules studied, wavelengths, λ, were converted to wavenumbers, ν, and emission intensities at each wavenumber were multiplied first by λ^2 and then by λ^3. The first factor compensates for the fixed-wavelength resolution used to measure the spectra and the second for the increased rate of emission with increased energy of emission.[50] Integrated intensity (A), average position ($\bar{\nu}$) and line width (Δ) were then calculated from the spectra $I(\nu)$ as

$$A = \int I(\nu) d\nu$$

$$\bar{\nu} = \frac{1}{A} \int I(\nu) \nu d\nu$$

$$\Delta = \left[\frac{1}{A} \int I(\nu)(\nu - \bar{\nu})^2 d\nu \right]^{0.5}$$

using the trapezoidal rule.

The remaining transformations performed to generate Figures 7-2, 7-3, and 7-4 are as follows. Each sample's integrated intensity, *A*, underwent separate corrections to address variations in sample handling across wells and across plates. *A* was divided first by OD_{600} and second by the average OD_{600}-corrected *A* for the four GFP-S65T controls in each plate. On the one occasion that a GFP-S65T control was unusually dim, its *A* was not included in calculating the average. The positions of GFP-S65T controls were found to vary considerably more from plate to plate (standard deviation of 18 cm^{-1}) than from well to well within the same plate (2.4 cm^{-1}). For this reason we have reported each sample's average position relative to the average of the positions of the four GFP-S65T controls in its plate. Accordingly we report 2.4 cm^{-1} as the error in the position measurements reported in Figures 7-3 and 7-4. Only samples with $OD_{600} > 0.1$ were used for making Figures 7-2, 7-3, and 7-4.

Errors reported in Figure 7-2 were estimated by the following bootstrap method. For each library, the fraction of functional samples was first recalculated for one hundred unique subsets of samples generated by selecting two-thirds of all samples at random. Error was then estimated as the standard deviation of these calculations. A potential source of systematic error in these experiments is primer synthesis. It is possible that the nucleotides that are mixed to introduce degeneracy are not incorporated with equal

frequency even if they are initially present at equal concentrations in solution. This possibility raises the concern that increasing levels of preservation of function may be largely determined by stronger biases in nucleotide incorporation towards the sequence of GFP-S65T. However, this cannot be the case since the trend towards increasing diversity of function with increasing preservation of function indicates that there is considerable sequence diversity in the libraries with the highest levels of preservation of function.

Statistics represented by the box plots in Figure 7-3 were calculated using Kaleidagraph v3.6 (Synergy Software). Since some quantitative estimates of diversity of function are based on relatively few fluorescent samples (e.g., extremes of function for the epPCR library or dispersion of function for the Random library), one might expect some variability in these estimates if this test were repeated. However, the overall qualitative trends are expected to be quite robust since we have over-sampled each designed library.

Acknowledgements

This research was supported by the Howard Hughes Medical Institute and the Army Research Office. We thank Patrick Daugherty for providing many of the primers used to assemble GFP-S65T and a pBAD-derived vector engineered with *Sfi*I recognition sequences; Christina Smolke for the use of her plate reader; Marco Mena, Michelle Meyer and Frances Arnold for advice in designing this project; and Marie Ary for useful comments on the manuscript. T.P.T. was supported by NIH Grant F32-GM07438. C.L.V. was supported by an NSF Graduate Research Fellowship.

References

1. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82–87.

2. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513.

3. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold. Des.* **7**, R105–R109.

4. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368.

5. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science* **304**, 1967–1971.

6. Saven, J. G. & Wolynes, P. G. (1997). Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. *J. Phys. Chem. B* **101**, 8375–8389.

7. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. G. (2001). Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. USA* **98**, 3778–3783.

8. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558.

9. Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A. & Dahiyat, B. I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. USA* **99**, 15926–15931.

10. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002). Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. *Protein Sci.* **11**, 2804–2813.

11. Moore, G. L. & Maranas, C. D. (2003). Identifying residue-residue clashes in protein hybrids by using a second-order mean-field approach. *Proc. Natl. Acad. Sci. USA* **100**, 5091–5096.

12. Endelman, J. B., Silberg, J. J., Wang, Z. G. & Arnold, F. H. (2004). Site-directed protein recombination as a shortest-path problem. *Protein Eng. Des. Sel.* **17**, 589–594.

13. Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H. & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518.

14. Ness, J. E., Kim, S., Gottman, A., Pak, R., Krebber, A., Borchert, T. V., Govindarajan, S., Mundorff, E. C. & Minshull, J. (2002). Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat. Biotechnol.* **20**, 1251–1255.

15. Coco, W. M., Encell, L. P., Levinson, W. E., Crist, M. J., Loomis, A. K., Licato, L. L., Arensdorf, J. J., Sica, N., Pienkos, P. T. & Monticello, D. J. (2002). Growth factor engineering by degenerate homoduplex gene family recombination. *Nat. Biotechnol.* **20**, 1246–1250.

16. Hogrefe, H. H., Cline, J., Youngblood, G. L. & Allen, R. M. (2002). Creating randomized amino acid libraries with the QuikChange Multi Site-Directed Mutagenesis Kit. *Biotechniques* **33**, 1158–1160, 1162, 1164–1165.

17. Hiraga, K. & Arnold, F. H. (2003). General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330**, 287–296.

18. Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z. G. & Arnold, F. H. (2003). Library analysis of SCHEMA-guided protein recombination. *Protein Sci.* **12**, 1686–1693.

19. Otey, C. R., Silberg, J. J., Voigt, C. A., Endelman, J. B., Bandara, G. & Arnold, F. H. (2004). Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach. *Chem. Biol.* **11**, 309–318.

20. Heim, R., Cubitt, A. B. & Tsien, R. Y. (1995). Improved green fluorescence. *Nature* **373**, 663–664.

21. Jain, R. K. & Ranganathan, R. (2004). Local complexity of amino acid interactions in a protein core. *Proc. Natl. Acad. Sci. USA* **101**, 111–116.

22. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–88.

23. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.* **240**, 421–433.

24. Shafikhani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. (1997). Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. *Biotechniques* **23**, 304–310.

25. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. USA* **102**, 606–611.

26. Shagin, D. A., Barsova, E. V., Yanushevich, Y. G., Fradkov, A. F., Lukyanov, K. A., Labas, Y. A., Semenova, T. N., Ugalde, J. A., Meyers, A., Nunez, J. M., Widder, E. A., Lukyanov, S. A. & Matz, M. V. (2004). GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. *Mol. Biol. Evol.* **21**, 841–850.

27. Patrick, W. M., Firth, A. E. & Blackburn, J. M. (2003). User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng.* **16**, 451–457.

28. Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116.

29. Wang, L., Jackson, W. C., Steinbach, P. A. & Tsien, R. Y. (2004). Evolution of new nonantibody proteins via iterative somatic hypermutation. *Proc. Natl. Acad. Sci. USA* **101**, 16745–16749.

30. Cormack, B. P., Valdivia, R. H. & Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38.

31. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **3**, 253–258.

32. Heim, R. & Tsien, R. Y. (1996). Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr. Biol.* **6**, 178–182.

33. Ormo, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y. & Remington, S. J. (1996). Crystal structure of the Aequorea victoria green fluorescent protein. *Science* **273**, 1392–1395.

34. Cherry, J. R. & Fidantsef, A. L. (2003). Directed evolution of industrial enzymes: an update. *Curr. Opin. Biotechnol.* **14**, 438–443.

35. Bokman, S. H. & Ward, W. W. (1981). Renaturation of Aequorea green-fluorescent protein. *Biochem. Biophys. Res. Commun.* **101**, 1372–1380.

36. Noivirt, O., Eisenstein, M. & Horovitz, A. (2005). Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng. Des. Sel.* **18**, 247–253.

37. Fodor, A. A. & Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211–221.

38. Lovell, S. C., Davis, I. W., Adrendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Structure validation by C alpha geometry: phi, psi and C beta deviation. *Proteins* **50**, 437–450.

39. Mayo, S. L., Olafson, B. D. & Goddard, W. A. (1990). Dreiding — a Generic Force-Field for Molecular Simulations. *J. Phys. Chem.* **94**, 8897–8909.

40. Helms, V., Winstead, C. & Langhoff, P. W. (2000). Low-lying electronic excitations of the green fluorescent protein chromophore. *Theochem-J. Mol. Struct.* **506**, 179–189.

41. Sitkoff, D., Sharp, K. A. & Honig, B. (1994). Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **98**, 1978–1988.

42. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**, 10172–10177.

43. Connolly, M. L. (1983). Solvent-Accessible Surfaces of Proteins and Nucleic-Acids. *Science* **221**, 709–713.

44. Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.

45. Desmet, J., Demaeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539–542.

46. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J. Comput. Chem.* **24**, 232–243.

47. Stemmer, W. P. C., Crameri, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. (1995). Single-Step Assembly of a Gene and Entire Plasmid from Large Numbers of Oligodeoxyribonucleotides. *Gene* **164**, 49–53.

48. Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. (1995). Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose P-Bad Promoter. *J. Bacteriol.* **177**, 4121–4130.

49. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K. & Pease, L. R. (1989). Site-Directed Mutagenesis by Overlap Extension Using the Polymerase Chain-Reaction. *Gene* **77**, 51–59.

50. Lakowicz, J. R. (1999). *Principles of Fluorescence Spectroscopy*. Second edit., Springer, New York.

Table 7-1: Library designs

Pos	$DBIS^{ORBIT}$	$DBIS^{ORBIT}4^4$	C^{ORBIT}	$SCMF^{ORBIT}32^2$	C^{MSA}	SE/C^{MSA}	Random
57	W	W	W	W	W	W	W
58	PA	PAST	PT	all	P	PH	PQ
59	TS	T	TS	T	TI	T	TN
60	L	L	L	L	L	L	L
61	VL	VALS	VL	V	V	VI	VD
62	TA	TAGS	TA	T	TA	TA	TN
63	T	T	TA	T	TA	TA	T
64	F	F	F	F	FL	FL	F
65	TA	T	TA	T	TS	TS	TK
67	G	G	G	G	G	G	G
68	VA	V	V	V	VF	VF	VM
69	QL	QELV	QL	Q	QR	QR	QE
70	C	C	C	all	C	C	C
71	FL	F	FL	F	FY	F	FY
72	SA	S	SA	S	SA	SA	SI

The first amino acid listed at each position is that of GFP-S65T. Underlined amino acids are mutations designed as described in Methods.

Table 7-2: Analysis of mutations

mut.	*δ(peak position)*	*δ(peak width)*	*δ(Stoke's shift)*	*δ(T_m)*	*n*[*]
P58A	3.5 ± 62.6	5.9 ± 20.6	-93.0 ± 315.3	-3.1 ± 2.4	7
T59S	21.6 ± 33.5	4.2 ± 10.5	62.4 ± 191.6	1.0 ± 3.3	13
V61L	14.2 ± 48.6	-3.3 ± 6.3	94.6 ± 285.6	-1.1 ± 0.7	4
T62A	-22.7 ± 34.6	8.3 ± 15.2	154.5 ± 108.6	-6.3 ±2.5	10
T65A	128.9 ± 28.0	28.7 ± 17.3	103.6 ± 157.3	1.7 ± 2.9	12
V68A	-30.5 ± 50.3	12.7 ± 7.1	-149.2 ± 101.1	-2.1 ± 3.1	5
Q69L	65.2 ± 54.5	30.8 ± 15.3	91.6 ± 257.5	7.1 ± 3.9	11
F71L	28.2 ± 42.1	11.5 ± 16.0	100.9 ± 260.1	0.3 ± 3.1	7
S72A	-85.9 ± 58.9	-28.8 ± 18.9	-326.7 ± 270.9	-0.7 ±3.5	11

[*] *n* is the number of mutational backgrounds in which a given mutation was sampled in the sequenced library members.

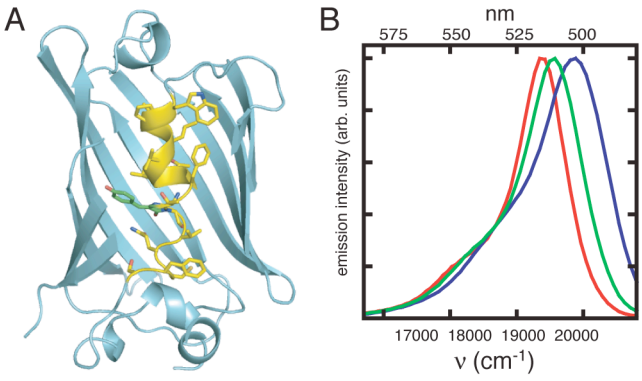


Figure 7-1. Structure of GFP-S65T and spectra of variants. (A) The front side of this cylindrical protein has been clipped to spotlight residues 57–72 in its core. Side chain atoms for targeted positions 57–65 and 67–72 are illustrated in CPK colors with carbon in yellow. The chromophore of GFP is shown in CPK colors with carbon in green. This figure was composed from a 1.45 Å-resolution structure of GFP containing the S65T and Q80R mutations (PDB code 1q4a).[21] (B) Extremes of function. Of the 11,575 spectra measured, 701 were at least one-half as bright as spectra of cultures known to express GFP-S65T. Of these, the red-most spectrum was sampled from the epPCR library (red), and the blue-most spectrum was sampled from the C^{ORBIT} library (blue). The spectrum of a culture expressing GFP-S65T is shown in green. The three spectra have been normalized to the same peak intensity.

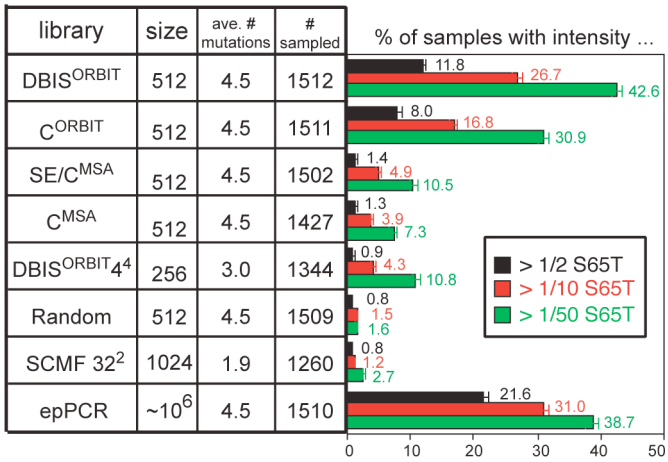


Figure 7-2. Preservation of function. A sample is variously defined as being functional if its emission intensity is at least one-half (black), one-tenth (red), or one-fiftieth (green) the intensity of cultures expressing GFP-S65T. Designed libraries are listed from top to bottom according to preservation of function calculated by the most exclusive definition. The theoretical library size, the average number of mutations, and the number of clones sampled are listed for each library.

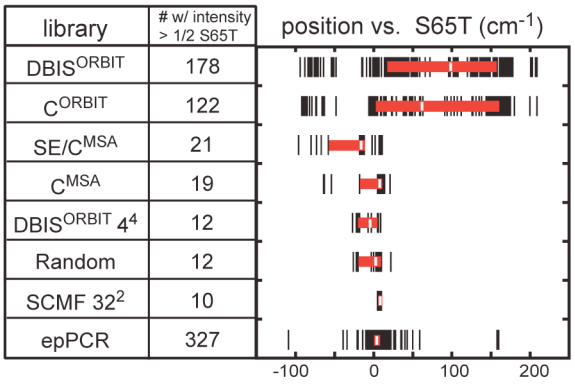


Figure 7-3. Diversity of function. Considering only those spectra with at least one-half the intensity of cultures expressing GFP-S65T, this plot illustrates the set of colors sampled from each library (black marks), the median of each set (white bar), and the first and third quartiles (red box). Positions are calculated relative to GFP-S65T standards as described in the Methods section. Designed libraries are listed from top to bottom according to preservation of function calculated by the most exclusive definition of function.

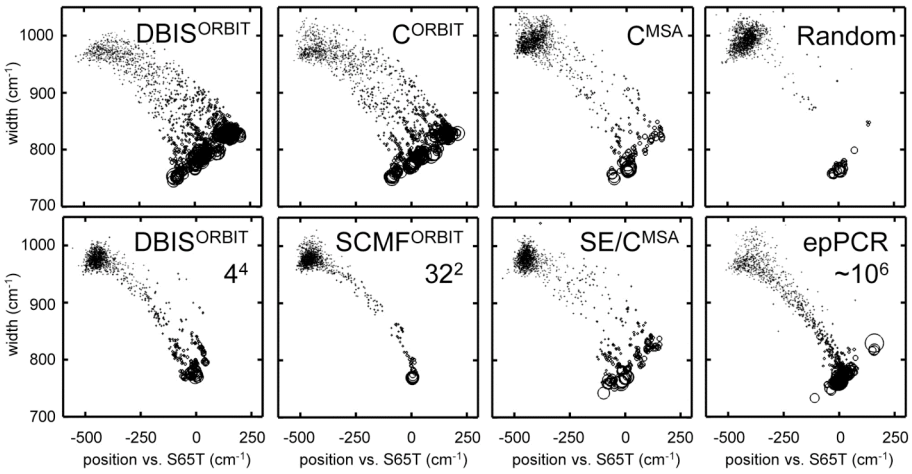


Figure 7-4. Preservation and diversity of function. The width of each spectrum sampled is plotted against its average position with a circle of area proportional to its integrated intensity. The brightest cultures (largest circles) were found to emit three orders of magnitude more light than the darkest cultures, such that the latter have the appearance of dots on these plots. Most dots cluster in the upper-left-hand corner of each plot, where the intrinsic fluorescence spectrum of these cultures is found. The brightest cultures, including those expressing GFP-S65T, cluster in the lower-right-hand corner of each plot. These plots illustrate that the seven designed libraries cluster into four performance categories. The $DBIS^{ORBIT}$ and C^{ORBIT} libraries performed best in two respects: first, as conveyed by the relative sparesensess of dots in the upper-left-hand corners of these plots, these libraries had by far the smallest fractions of non-functional variants sampled; second, the functional variants in these libraries can be seen to be distributed rather evenly across a relatively large range of positions. In contrast the functional variants in the SE/C^{MSA} and C^{MSA} libraries sample a somewhat smaller range of positions, and those at the extremes of this range tend to be more weakly fluorescent than those in the middle. The $DBIS^{ORBIT}$ 4^4 and Random libraries form the third performance category since most, but not all of the functional variants in these libraries closely resemble GFP-S65T in color. The $SCMF^{ORBIT}$ 32^2 library forms the last category, since its few functional variants exhibit the least variation in color. It is interesting to note that for most libraries we find the brightest cultures exhibit the same linear correlation between spectral width and average position. We have investigated the physical mechanisms that may be responsible for this correlation and present them elsewhere Treynor *et al.* (in preparation).

A

```
          --60--------70-
S65T      PTLVTTFTYGVQCFS   n
          -S---------L--A   1
          -S-----A--A----   1
          -------------A   11
          ------A--A---A   4
          --------------   9
          AS----------LA   2
          -S-L---A--A---A   1
          -S-----A---L---   3
          -S--A--A------A   3
          -S-----A-------   9
          ------A------A   10
          -S------------   3
          ------A--A----   3
          AS-----A------A   3
          ----A---------A   1
          AS-L---A------A   2
          -S--A--A---L--A   1
          ----A--A------A   1
          -S-L---A-----LA   1
          ---------A---A   2
          -S-L---A-------   2
          -S------------A   1
          ----------L--A   6
          ----A--A---L--A   2
          -S-----A-----L-   1
          -S-----A---L-LA   1
          AS--A--A------A   1
          -S--A---------A   1
          ---L----------A   1
          -------A-----LA   1
          A------------A   1
          A------A--A----   1
          AS-----A-----LA   1
          ------------LA   1
          A------A--A---A   1
          ------A---L--A   7
          -S--A--A-------   1
          ----A--A--A---A   1
          -------A---L-LA   1
          -S-----A------A   1
          -S-----A-----LA   1
          -S-----A---L--A   1
          -------A-------   5
          ----------L---   5
          ------A---L---   5
          A---A--A--A---A   4
```

B

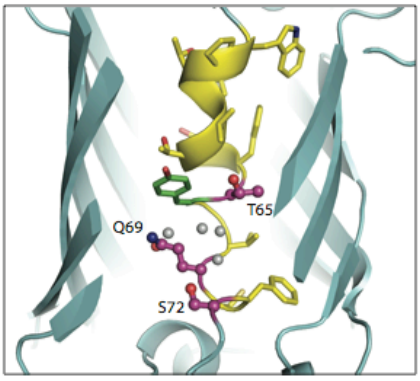


Figure 7-5. Mutational analysis of GFP-S65T variants from DBIS^{ORBIT} library. (A) Sequences of bright GFP-S65T variants. For clarity, only the residues (58–72) that were mutated in the designed library is shown, with the parent sequence (S65T) at the top. The number of wells for which that sequence was sampled (*n*) is also shown. A dash in the sequence indicates the WT amino acid identity. (B) A close-up of the chromophore region, showing the residues that cause significant shifts in emission peak position (magenta, balls and sticks). Several ordered waters in the crystal structure are shown as grey spheres.

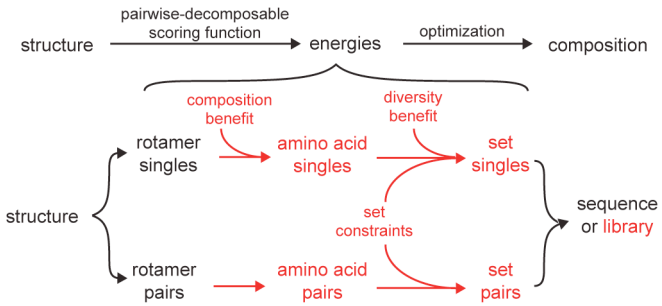
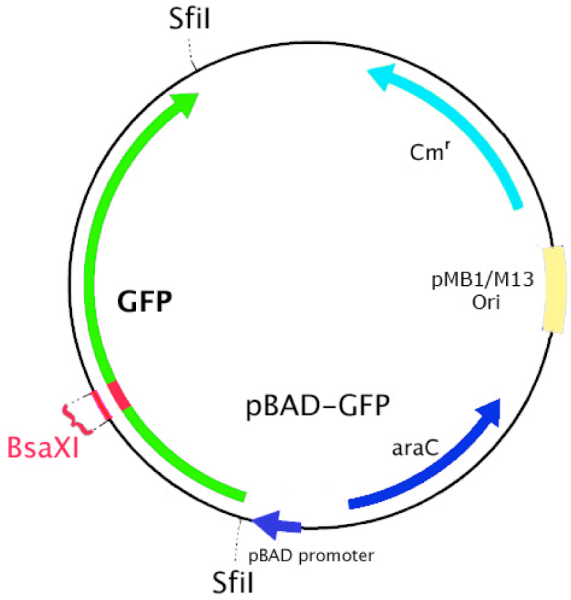


Figure 7-6. The DBIS algorithm. The flow chart at top illustrates the core procedure shared by many algorithms used for the structure-based computational design of either proteins or combinatorial libraries. The flow chart at bottom illustrates the main components of the generalized DBIS algorithm. If the components shown in red were eliminated, the remaining components would be sufficient to design a single protein instead of a library.




```
5′–     TAATGCATGACCTCGACTCCGCTGAGCAGA  – 3′
3′– GGTATTACGTACTGGAGCTGAGGCGACTCG      – 5′
```

Figure 7-7. Vector map and *Bsa*XI site. Top: map of vector derived from pBAD18-Cm.[48] Two unique *Sfi*I recognition sequences were added after the pBAD promoter for unidirectional gene insertion. The *Bsa*XI recognition sequence and a small restriction fragment were substituted for nucleotides corresponding to positions 57–72 as described in Methods. Bottom: small restriction fragment produced by digestion of this vector with *Bsa*XI. Nucleotides in red are the *Bsa*XI recognition sequence. Underlined nucleotides are a *Blp*I recognition sequence. Overhanging ends in bold correspond to positions 56 and 73.

Appendix A

Double mutant cycle analysis of an ion pair on the surface of protein G

Abstract

The role of ion pairs and salt bridges on the surfaces of proteins is unclear. Ion pairs have been identified in which interactions have either negligible or significant contribution to stability. It has been hypothesized that for cases in which there is no net contribution to stability or weak interaction energy, the electrostatic interaction may aid in fold specificity. Here we use double mutant cycle analysis to evaluate the interaction energy of an ion pair on the beta sheet surface of the B1 domain streptococcal protein G. The interaction energy between Lys4 and Glu15 is found to be favorable by 0.78 ± 0.29 kcal mol^{-1} at 75°C. Mutation of both residues simultaneously indicates that they contribute little to the stability of the native state relative to the stability of the double mutant K4T/E15T.

Introduction

Electrostatic interactions can play an important role in stabilizing the folded state of a protein.[1-3] Thermophilic proteins contain an increased number of salt bridges and ion pairs over their mesophilic counterparts.[1] Experimental analysis has quantified the role by which electrostatic interactions can stabilize the native state.[2,4] On the other hand, equivalent mutagenesis studies in different systems have revealed that these interactions can contribute little energy to the stabilization of the native fold.[4-7] This has led to the proposal that these interactions may be more important for fold specificity than thermodynamic stability.[6,8] The environment of the ion pair most likely plays a role in the degree to which the interaction stabilizes the native state. For instance, solvent-exposed interactions may have negligible contribution to the free energy of folding because their formation comes at an entropic cost and also because the interaction is screened by counterions and polar solvent. However, ion pairs that are shielded from solvent have a desolvation penalty associated with the buried charges.[9] The contribution to stability of salt bridges and ion pairs in some proteins have been explained by examining long-range electrostatic interactions between the interacting pair and the rest of the protein.[4]

Here we evaluate the interaction energy of the ion pair formed by Lys4 and Glu15 on surface of the B1 domain of protein G (GB1). GB1 is a model system for design due to its compact structure, large unfavorable free energy of unfolding (ΔG_u), and high thermal stability.[10] Residues 4 and 15 in the WT protein form an ion pair between anti-parallel strands 1 and 2 on the beta-sheet surface (Figure A-1A). Here we refer to ion pairs as close-range interactions between oppositely charged groups that are not

hydrogen-bonded to each other whereas salt bridges refer to hydrogen-bonded interactions. The ORBIT energy function does not recognize the interaction between Lys4 and Glu15 in the crystal structure of GB1 as a hydrogen bond. The nitrogen atom of the Lys sidechain is within 4 Å of the Glu carboxylic oxygen atoms. It has proven difficult to increase the thermostability of GB1 by optimizing the amino acid sequence of the beta-sheet surface. This has not been the case for the core residues of GB1.[11] The amino sidechains on the beta-sheet surface may be optimized for stability.

The standard method for evaluating the energy of interaction of two residues is the double mutant cycle.[4,12] This method separates the energetic cost of mutating each interacting residue separately from the energetic cost of mutating them simultaneous. If the sidechains have no interaction, the result of mutating one should be independent of the identity of the other. Double mutant cycle analysis assumes that the native state fold is not perturbed in the mutant sequences.[4] Figure A-1B shows the mutational cycle that is investigated in this study. We hypothesized that the packing interaction between Lys4, Glu15 and Ile6 might lead to cooperativity in the network such that the entropic cost of forming the Lys4-Glu15 ion pair is partially compensated by the packing interaction of both residues with Ile6. Such a synergistic effect has been observed in other networks of interacting residues.[13,14]

Methods

GB1 variants were generated using inverse PCR. Proteins were expressed in *E. coli* BL21-DE3 cells using IPTG induction and a pet11a expression plasmid. Cells were lysed by freeze-thaw cycling. Lysates were combined 1:1 with acetonitrile, clarified by

centrifugation, and purified by reverse phase HPLC using a C8 column with a linear 0.1% TFA/acetonitrile gradient. Eluted peaks were rotavapped to remove acetonitrile, flash frozen and lyophilized. Protein masses were confirmed by mass spectrometry. CD data was collected on an Aviv 62DS spectrometer equipped with a thermoelectric unit and an autotitrator. The protein samples were dissolved in 50 mM sodium phosphate buffer at pH 6.5. Thermal denaturation were collected by monitoring the CD signal at 218 nm using 1°C temperature steps from 1°C to 99°C with 2 minutes of equilibration time and 30 seconds of signal averaging time at each step. Guandinium hydrochloride concentration was determined by refractometry. Titrations were carried out in 0.2 M concentration steps with 10 minutes of stirring and 1 minute of signal averaging time at each step.

Results and Discussion

In order to maintain the fold of the beta-sheet surface, each residue was mutated to Thr, an amino acid with high propensity to form beta-sheet secondary structure.[15,16] It should be noted that the energy of interaction derived for a double mutant cycle is relative to the interaction energy of the mutant residues.[4] Therefore to claim that the true interaction between sidechains 4 and 15 has been measured, it must be assumed that Thr sidechains at those same positions do not interact.

All four variants had a CD signal characteristic of the WT protein (data not shown). The stability of each variant was assessed using thermal and chemical denaturation (Figure A-2). Free energies of unfolding, ΔG_u, were determined assuming a two-state transition with a temperature-independent heat capacity chage, ΔC_p (621 cal

mol^{-1} K^{-1}).[7,15] Briefly, the relationships between equilibrium constant K, free energy ΔG, enthalpy ΔH, and entropy

$$\Delta G = -RT \ln K \tag{1a}$$

$$\Delta G(T) = \Delta H_m \left(1 - \frac{T}{T_m}\right) + \Delta C_p \left[T - T_m - T \ln\left(\frac{T}{T_m}\right)\right] \tag{1b}$$

$$K = \frac{\theta_n - \theta}{\theta - \theta_u} \tag{1c}$$

can be combined in an expression for the spectroscopic signal, θ, in terms of the temperature T, thermal denaturation temperature T_m, gas constant R, enthalpy of denaturation at T_m, ΔH_m, and ΔC_p:

$$\theta = \theta_u + \frac{\theta_f - \theta_u}{1 + \exp\left[\frac{-\Delta H_m}{R}\left(\frac{1}{T} - \frac{1}{T_m}\right) + \frac{\Delta C_p}{R}\left(\left(\frac{T_m}{T} - 1\right) + \ln\left(\frac{T}{T_m}\right)\right)\right]}. \tag{2}$$

Melting temperatures and enthalpies of unfolding derived from Equation 2 were substituted back into Equation 1b to get $\Delta G(T)$. Chemical denaturation data was analyzed using the linear extrapolation method to obtain the free energy of unfolding at zero denaturant concentration.[17] Thermodynamic data is given in Table A-1.

Free energies of unfolding were used to calculate the free energy of interaction, $\Delta\Delta G^{KE}$, between Lys4 and Glu15

$$\Delta\Delta G^{KE} = (\Delta G^{KE} - \Delta G^{TT}) - [(\Delta G^{KT} - \Delta G^{TT}) + (\Delta G^{TE} - \Delta G^{TT})]. \tag{3}$$

The error in the free energy of interaction was obtained by propagating the error in Equations 1b and 3

$$\boldsymbol{\Delta}^2(\Delta G(T)) = \left[\frac{\delta(\Delta G(T))}{\delta(\Delta H_m)}\right]^2 \boldsymbol{\Delta}^2(\Delta H_m) + \left[\frac{\delta(\Delta G(T))}{\delta(T_m)}\right]^2 \boldsymbol{\Delta}^2(T_m) \tag{4a}$$

$$\boldsymbol{\Delta}^2(\Delta G(T)) = \left[(1 - \frac{T}{T_m}\right]^2 \boldsymbol{\Delta}^2(\Delta H_m) + \left[\Delta H_m\left(\frac{T}{T_m^2}\right) - \Delta C_p + \frac{T}{T_m}\right]^2 \boldsymbol{\Delta}^2(T_m) \tag{4b}$$

$$\boldsymbol{\Delta}^2(\Delta\Delta G^{KE}) = \boldsymbol{\Delta}^2(\Delta G^{KE}) + \boldsymbol{\Delta}^2(\Delta G^{TT}) + \boldsymbol{\Delta}^2(\Delta G^{KT}) + \boldsymbol{\Delta}^2(\Delta G^{TE}) \quad (4c)$$

where the boldface $\boldsymbol{\Delta}$ symbols indicate error in the corresponding parameter. For ΔH_m and T_m, the error results from the non-linear fit of Equation 2 to the experimental data. From Equations 3 and 4, we obtain an energy of interaction for Lys4 and Glu15 of 0.55 ± 0.56 kcal mol^{-1} at 25 °C. To reduce uncertainty in ΔG_u by minimizing the extrapolation from the T_m in Equation 1b, the free energy of interaction at 75°C was calculated to be 0.78 ± 0.29 kcal mol^{-1}. Therefore, the Lys4-Glu15 ion pair has a negligible interaction energy, within the error of the analysis, at 25 °C and a favorable interaction energy at 75°C.

The contribution to the overall stability of GB1 by this ion pair, relative to the K4T/E15T mutant, seems to be insignificant: the free energy of unfolding of WT, 8.0 kcal mol^{-1}, is less than that of the K4T/E15T double mutant, 9.0 kcal mol^{-1}. Makhatadze et al. have suggested that medium and long-range electrostatics can effect the contribution of an ion pair to stability.[4] In the case of GB1, the rest of the protein, other than residues 4 and 15, create a negative electrostatic potential at both sites 4 and 15 (data not shown), indicating that the reverse ion pair (Glu4-Lys15) may have similar properties as the WT ion pair. This negligible contribution to stability measured in this double mutant cycle could be a result of the higher beta-strand propensity of Thr over the WT amino acids. Assessment of the interaction energy with different reference states in the double mutant cycle could shed light on this effect. However, the integrity of the beta-sheet should be assessed carefully for any Ala mutants at residues 4 and 15.

Although we cannot directly compare the free energies of unfolding obtained from thermal and chemical denaturation, it is interesting to note that the results obtained

using the two different methods show different relative stabilities for the variants. In fact, the $\Delta\Delta G^{KE}$ obtained from fits to the chemical denaturation curves has the opposite sign (-0.56 ± 0.18 kcal mol^{-1}) as the $\Delta\Delta G^{KE}$ derived from the thermal denaturation fits. This underscores the fact that charged chemical denaturants can confound the analysis of electrostatic interactions.[6] It would be desirable to follow up on the data presented here by measuring $\Delta\Delta G^{KE}$ using urea as a denaturant. However, WT GB1 does not unfold completely at high concentrations (~ 10 M) of urea.

Our data suggests that Ile6 could lead to a favorable interaction energy between Lys4 and Glu15. These results is are not completely consistent with the data collected by Lassila et al. for an engineered electrostatic triad on the surface of protein G.[7] To test definitively whether the residues 4-6-15 form an interaction network, the data presented here could be extended to a triple mutant cycle in which Ile6 is mutated to Val or Ala to reduce packing interactions.

References

1. Szilagyi, A. & Zavodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **8**, 493–504.

2. Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V. & Makhatadze, G. I. (2006). Protein stability and surface electrostatics: A charged relationship. *Biochem.* **45**, 2761–2766.

3. Makhatadze, G. I., Loladze, V. V., Gribenko, A. V. & Lopez, M. M. (2004). Mechanism of thermostabilization in a designed cold shock protein with optimized surface electrostatic interactions. *J. Mol. Biol.* **336**, 929–942.

4. Makhatadze, G. I., Loladze, V. V., Ermolenko, D. N., Chen, X. F. & Thomas, S. T. (2003). Contribution of surface salt bridges to protein stability: Guidelines for protein engineering. *J. Mol. Biol.* **327**, 1135–1148.

5. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.

6. Strop, P. & Mayo, S. L. (2000). Contribution of surface salt bridges to protein stability. *Biochem.* **39**, 1251–1255.

7. Lassila, K. S., Datta, D. & Mayo, S. L. (2002). Evaluation of the energetic contribution of an ionic network to beta-sheet stability. *Protein Sci.* **11**, 688–690.

8. Merkel, J. S., Sturtevant, J. M. & Regan, L. (1999). Sidechain interactions in parallel beta sheets: the energetics of cross-strand pairings. *Structure* **7**, 1333–1343.

9. Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **3**, 211–226.

10. Alexander, P., Fahnestock, S., Lee, T., Orban, J. & Bryan, P. (1992). Thermodynamic Analysis of the Folding of the Streptococcal Protein-G Igg-Binding Domains B1 and B2 — Why Small Proteins Tend to Have High Denaturation Temperatures. *Biochem.* **31**, 3597–3603.

11. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475.

12. Serrano, L., Horovitz, A., Avron, B., Bycroft, M. & Fersht, A. R. (1990). Estimating the Contribution of Engineered Surface Electrostatic Interactions to Protein Stability by Using Double-Mutant Cycles. *Biochem.* **29**, 9343–9352.

13. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A. R. (1990). Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.* **216**, 1031–1044.

14. Spek, E. J., Bui, A. H., Lu, M. & Kallenbach, N. R. (1998). Surface salt bridges stabilize the GCN4 leucine zipper. *Protein Sci.* **7**, 2431–2437.

15. Minor, D. L. & Kim, P. S. (1994). Measurement of the Beta-Sheet-Forming Propensities of Amino-Acids. *Nature* **367**, 660–663.

16. Street, A. G. & Mayo, S. L. (1999). Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9074–9076.

17. Pace, C. N. & Shaw, K. L. (2000). Linear extrapolation method of analyzing solvent denaturation curves. *Proteins* **Suppl. 4**, 1–7.

Table A-1: Thermodynamic data for GB1 variants

variant	T_m (°C)	ΔH_m (kcal mol^{-1})	$\Delta G(25°C)^{\S}$ (kcal mol^{-1})	$\Delta G(75°C)^{\S}$ (kcal mol^{-1})	$\Delta G(25°C)$ $LEM^{\dagger}$ (kcal mol^{-1})
WT	83.5 ± 0.3	68.0 ± 1.7	8.0 ± 0.3	1.6 ± 0.2	4.85 ± 0.02
K4T	78.8 ± 0.2	70.0 ± 1.4	8.0 ± 0.2	0.74 ± 0.11	5.93 ± 0.08
E15T	82.7 ± 0.3	71.1 ± 2.1	8.5 ± 0.4	1.5 ± 0.2	5.43 ± 0.15
K4T/E15T	82.1 ± 0.2	74.8 ± 1.2	9.0 ± 0.2	1.5 ± 0.1	5.95 ± 0.06
		$\Delta\Delta G^{KE}$:	0.55 ± 0.56	0.78 ± 0.29	-0.56 ± 0.18

* T_m and ΔH_m derived from from two-state fit of thermal denaturation curves (Equation 2)

§ ΔG calculated using Gibbs-Helmholtz equation (Equation 1b) and a constant ΔC_p = 621 cal mol^{-1}K^{-1}

† ΔG calculated using linear extrapolation method (LEM)

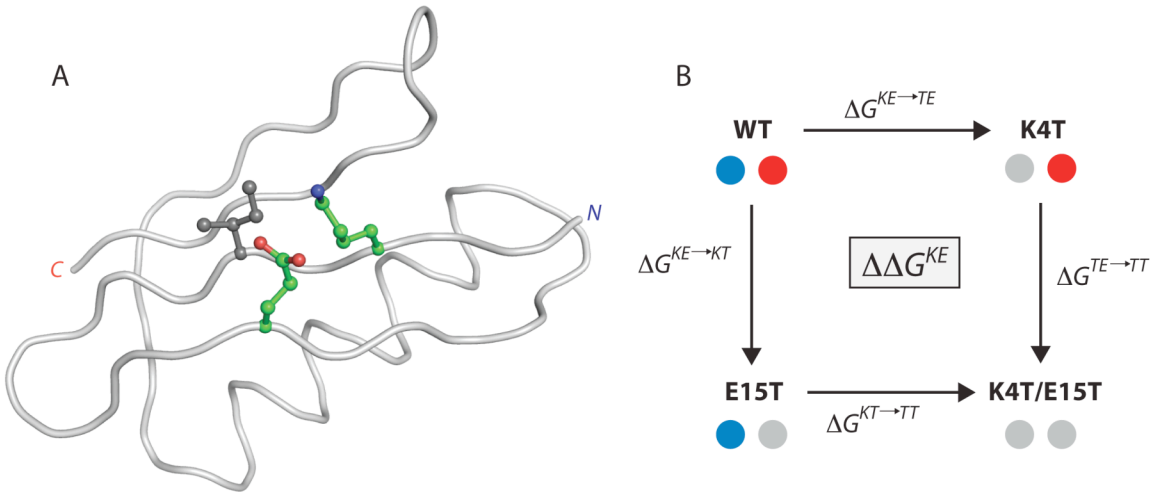


Figure A-1. The Lys4–Glu15 salt bridge in GB1. (A) Lys4, Ile6, Glu15 are shown in ball and stick representation (Ile6 in gray; Lys4,Glu15 in CPK colors). (B) The double mutant cycle examined in this study is shown with charged sidechains shown in colors and neutral sidechains shown in gray.

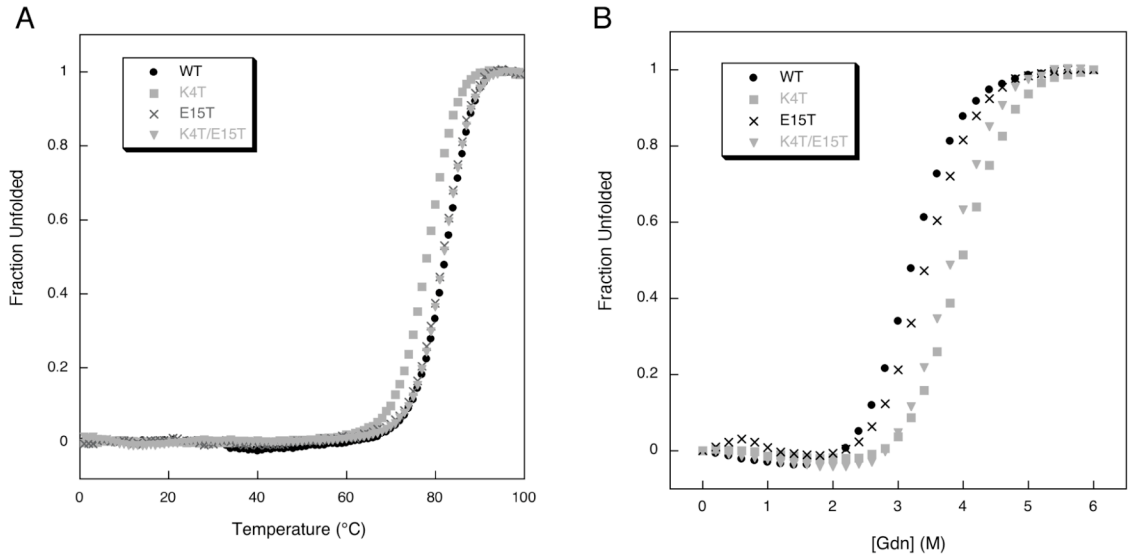


Figure A-2. CD data for GB1 variants. (A) Thermal denaturation curves for proteins in 50 mM sodium phosphate at pH 6.5. (B) Chemical denaturation in guanadinium chloride. Both plots were normalized using pre- and post-transition baselines from the raw data.

Appendix B

Evaluation of the Generalized Born model for computational protein design

Abstract

We have assessed the utility of the generalized Born (GB) model for use in protein design by looking at the model's accuracy for calculating rotamer self and pair energy terms. Energies from the GB model were also compared to values obtained using a finite difference Poisson-Boltzmann solver. Both implementations of the GB model studied here show promising one- and two-body decomposability. However, their accuracy is not significantly better than the accuracy of solvation and electrostatics models currently used in the ORBIT energy function.

Background

An accurate model of the aqueous environment is important for the design of well-folded, stable proteins.[1] Hydration of polar amino acids and burial of hydrophobic amino acids is the key determinant in folding and stability.[2] A general representation of free energy of solvation, ΔG_{solv}, is given by

$$\Delta G_{solv} = \Delta G_{np} + \Delta G_{np} \quad (1)$$

where ΔG_{np} is the free energy change due to placing a hypothetical nonpolar solute of the same volume as the actual solute into the solvent and ΔG_{pol} is the polar contribution to solvation, which is the free energy change from moving the solute's charge distribution from a non-polarizable environment to water.[3] Linear surface area-based scaling functions have been successful in expressing ΔG_{np},[4,5] but these models do not account for the factors contributing to ΔG_{pol}.[6]

A macroscopic continuum representation of water is often used to account for the polarizability of water (ε_{out}=80). Although the protein is modeled using a standard all-atom molecular mechanics representation, a dielectric constant is also assigned to the protein interior in order to capture the dielectric response, primarily electronic polarization, inside the protein molecule (ε_{in}=2-4).[7] The continuum representation of the solvent/solute system is a simplification, but it has worked well for many applications.[7-9] The Poisson equation is the fundamental equation governing the relationship between electrostatic potential (ϕ), dielectric response (ε) and charge distribution (ρ) in continuum systems:

$$\nabla \cdot [\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \quad . \quad (2)$$

The solution to this equation becomes an approximation when we assign a dielectric constant to the protein region, but Equation 2 is still considered the benchmark for evaluating the electrostatic energy of proteins and other macromolecules.[9] Various numerical methods have been implemented for solving the Poisson equation (or Poisson-Boltzmann, PB, for non-zero ionic strength). The finite difference method (FDPB), which involves distributing charges and dielectric constants over a grid and solving Equation 2 at each grid point, has been used extensively for biomolecular applications.[10] Due to the computational cost of using numerical methods, there is a great deal of interest in developing fast analytical methods for calculating electrostatic potential in a protein.

The generalized Born (GB) model is a fully analytical approximation of the Poisson equation.[11,12] It has been developed for use in molecular dynamics simulations due to the ease with which forces (first derivatives) can be calculated.[13-16] The GB equation, formulated by Still and coworkers[11], includes the Born energy[17] of each partial charge in the molecule as well as the screening energy of all charge-charge interactions:

$$\Delta G_{pol} = -166\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{ex}}\right)\sum_{i}^{N}\sum_{j}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i\alpha_j e^{(-r_{ij}^2/4\alpha_i\alpha_j)}}} \qquad (3)$$

where r_{ij} is the distance between charges q_i and q_j. The Born radius, α, is a parameter that captures the effective distance from the dielectric boundary of each charge or partial charge in the molecule. The accuracy of Born radii has proven to be essential in accurately calculating the electrostatic solvation energy, ΔG_{pol}.[18]

There are a number of methods for calculating Born radii. These methods include but are not limited to: the pairwise descreening approach (PDA),[19] the surface integral model,[20] the asymptotic approach (GBSA),[13,21] and the molecular volume approach.[22,23]

The computational speed and obvious generalization to design of the atomic pairwise approaches, like GBSA and PDA, make these models the most attractive for further study. Analytical calculation of Born radii is permitted by the assumption that the dielectric displacement is Coulombic in form, neglecting the reaction field component that would require iterative evaluation.[12,24] By integrating over the energy density of the Coulombic electrostatic field[25] and substituting back into the Born formula for the solvation energy of an ion, the Born radius can be expressed as a function of the van der Waals radius of the atom and the position of all other atoms in the protein.[9,12]

$$\frac{1}{\alpha_i} = \frac{1}{R_{vdW,i}} - \frac{1}{4\pi}\int_{solute, r>R_{vdW,i}} \frac{1}{r^4} dV \tag{4}$$

This simplification, termed the Coulomb Field Approximation (CFA), leads to an overestimation of the self-energy terms.[24] Correction factors have been proposed to account for the CFA,[23,26] and parameterization using FDPB energies are used to reproduce FDPB results within the limitations of the CFA.

The GBSA method approximates the amount of favorable charge/induced dipole interaction energy lost when a neutral atom (j) displaces the dielectric medium in proximity to a charge (i) to be V_j/r_{ij}^4, where r_{ij} is the distance between the charge and the atom and V_j is the volume of atom j.[21,27] The value V_j/r_{ij}^4 is an accurate evaluation of the Coulomb integral (Equation 4), as r_{ij} becomes large: it is equivalent to moving the $1/ r_{ij}^4$ out of the integral in Equation 4. Still and coworkers[21] propose scaling this term depending on the spatial relationship, e.g., bonded or non-bonded, between *i* and *j*. The original equation of Still and coworkers[21] was recast by Dominy and Brooks[13] to allow for a linear fitting in parameterization:

$$\Delta G_{pol,i} = \left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{ex}}\right)\left[\frac{1}{\lambda}\left(\frac{-166}{R_{vdW,i}}\right) + P_1\left(\frac{166}{R^2_{vdW,i}}\right) + \sum_j^{bond} \frac{P_2 V_j}{r_{ij}^4} + \sum_j^{angle} \frac{P_3 V_j}{r_{ij}^4} + \sum_j^{non-bond} \frac{P_4 V_j}{r_{ij}^4} CCF\right] \quad (5)$$

where P_1 - P_4 and λ are scaling factors that account for the varying inaccuracy of V_j/r_{ij}^4. The closest contact function (CCF) is a damping function that reduces the volume of atoms that are overlapping and it depends on a separate parameter, P_5. This method is effective at calculating molecular solvation energies and is computationally fast.[9,13]

The pair-wise descreening approximation (PDA) method for calculating Born radii, like the GBSA method, is based on the strategy of summing over atomic desolvation effects as estimated by integration of the Coulomb field energy density.[19] The difference between the PDA and GBSA methods is that, for PDA, the analytical solution for the Coulomb integral is evaluated over the spherical atomic volumes of atoms *j* surrounding atom *i*. The analytical evaluation of the descreening integral *H* is given in Reference 30

$$\alpha_i^{-1} = R_{vdW,i}^{-1} - \frac{1}{2}\sum_j H(r_{ij}, S_j(R_{vdW,j})) \quad . \quad (6)$$

In order to account for overlap among the descreening atoms *j*, these integrals are then scaled by factors, S_j, depending on the identity of atom *j*. This method has worked well for small molecules.[28] Case and coworkers suggested a modification to improve the PDA method's performance for large molecules.[29]

The problem with summing over atomic volumes (GBSA) or inter-atomic distances (PDA) in a continuum context, is that any region of the molecule not occupied by a solute atom will be filled by the high dielectric medium even though many such small cavities inside the protein are not large enough to accommodate a solvent molecule. These micro-dielectrics will cause a systematic underestimation in the Born radii of

deeply buried polar atoms. Since deeply buried atoms, for which the micro-dielectric problem will be most deleterious, generally contribute least to the overall solvation energy of the molecule, the error was considered acceptable. For large molecules, Case and coworkers proposed a scaling parameter on the second term of Equation 6 in order to remove the micro-dielectrics while maintaining accurate results for surface atoms.[29]

Below we report on the accuracy of the GB model, compared to FDPB calculations, as well as results showing the pairwise decomposability of two GB models. The simplified surfaces method, described in Chapter 3 of this thesis, is used to implement pairwise decomposable GB models. Inherent biases of the GB models are examined by comparing their atomic Born radii to ideal radii.

Methods

The GB method of Dominy and Brooks[13] (GBSA) was re-parameterized to give values for λ and P_1-P_5 that are consistent with the PARSE radii set.[30] The parameters were obtained through the same linear fitting to FDPB values as described by Dominy and Brooks.[13] Due to the desire for a parameter set to use in protein design calculations, an all-protein training set consisting of 22 single-chain high-resolution crystal structures was used. The pdb codes for these structures are: 1ajj, 2erl, 1ptq, 1pga, 1enh, 1vjw, 1igd, 1ptf, 1rge (chain A), 1rro, 2rhe, 1dhn, 1whi, 4fgf, 1tta (chain A), 194l, 2end, 2rn2, 2cpl, 3lzm, 1amm, and 1mrj. Hydrogens were added to the structures using MolProbity,[31] and no further modifications were made. Amino acid specific parameter sets were obtained by linear fitting with only the polar/charged atoms from the specific sidechain. Similarly,

a separate parameter set was obtained for backbone atoms only. The overlap volume correction of Still and coworkers[21] was used to partially account for overlapping atoms.

The original and modified GB methods of Case and coworkers[29,32] (GB-PDA) were used with the published parameter set[33] and Bondi radii.[34] The value of the packing correction factor for the original GB method of Case and coworkers (λ=1.4) was used. For all GB calculations, the protein dielectric, ε_{in}, was set to 4, and the solvent dielectric, ε_{ex}, was set to 80.

FDPB calculations were carried out using the DelPhi program[10] with a 0.5 Å grid spacing, 70% grid fill, zero ionic strength, $\varepsilon_{in} = 4$, $\varepsilon_{ex} = 80$, and the PARSE parameter set. Atomic FDPB solvation energies were calculated by placing a unit charge at the atom of interest and calculating the energy of the system with all other atoms neutral. DelPhi Born radii were obtained by substituting this atomic solvation energy into the Born equation and solving for ionic radius. The methods for calculating exact, one-body, and two-body screening and desolvation energies are described in Chapter 3 of this thesis.

Results and Discussion

Using the simplified surfaces method, the one- and two-body decomposability of the GBSA and GB-PDA methods are similar to that of the FDPB solver DelPhi (Figure B-1 and Table B-1). For the one-body approximation to backbone desolvation, the DelPhi approximation compares most favorably with the exact desolvation calculated with all sidechains present (Figure B-1C). As would be expected for one-body desolvation, the GBSA and GB-PDA methods overestimate the backbone desolvation when the one-body contributions are added together. Since one-body backbone

desolvation does not take into account an overlap between atoms on separate sidechains, some desolvation effects will be counted multiple times.

The one-body approximation to sidechain desolvation (graph not shown) is not well correlated with the exact sidechain desolvation for any of the methods. This is not surprising since one-body sidechain desolvation does not take into account desolvation of a sidechain by any other sidechains in the protein. The two-body sidechain desolvation results show a very different trend from the one-body backbone desolvation results. For the two-body decomposition, GB-PDA (Figure B-1E) is most consistent with the exact sidechain desolvation, while DelPhi (Figure B-1F) is the least decomposable. The DelPhi two-body sidechain desolvation shows a trend of underestimating the desolvation energy for sidechains that have a large desolvation penalty.

The GB-PDA method is the most decomposable of the three methods for the calculation of screened Coulombic energy (Figure B-2). For both sidechain/sidechain and sidechain/backbone values, the two-body decomposition for all methods show a tendency to underestimate the magnitude of these interaction energies. The reduced interaction energy is caused by over-screening of the interactions due to the reduced representation of the protein in the simplified surface model. This problem is more evident in the DelPhi two-body decomposition of sidechain/sidechain screening energy (Figure B-2F). The one-body approximation for sidechain/backbone screened Coulombic energy (data not shown) is nearly as accurate as the two-body decomposition, especially for the GB-PDA method.

The pairwise decomposability of a solvation model is a necessary but not sufficient qualification for the model to be used in design. The accuracy that one would

gain from using the GB model also determines whether the model is worth using. We have measured accuracy by comparing results with GB to those with DelPhi. In a comparison of FDPB solvers, DelPhi calculations with a 0.5 Å grid spacing were found to have comparable accuracy to other FDPB solvers with smaller grid spacings.[9] Since the majority of GB studies have looked at calculating molecular solvation energies, it is necessary to assess the accuracy of GB methods for calculating the energy terms that are used in protein design calculations and compare their accuracy with currently used methods for calculating electrostatic energy. All comparison with DelPhi energies in Figures B-3, B-4, B-5, and Table B-2 refer to the exact GB energy, not the one- or two-body decomposition.

For backbone and sidechain desolvation, the GBSA method overestimates the desolvation effect (Figures B-3A and B-3B), performing particularly poorly for backbone desolvation. The GB-PDA model gives accurate backbone desolvation but tends to underestimate sidechain desolvation. This trend in most likely a result of systematic underestimation of Born radii for atoms buried in folded proteins (Figure B-5E). The GB methods were compared with the LK solvent-exclusion model,[35] using an LK parameter set that had been tuned to reproduce PB energies (Marshall & Mayo, unpublished work). The sidechain desolvation RMSD and correlation values (Table B-2) show that all three models are similar. The accuracy of a more recently developed, modified GB-PDA method,[32] which uses a continuous expression for Born radii, (data not shown) is comparable to that of the original GB-PDA method shown in Figure B-3C.

For screened Coulombic energy, GB methods were compared with the distance-dependent dielectric (DDD) model currently used in the ORBIT energy function (Figure

B-4). For sidechain/backbone screened Coulombic energy, the GB models and the DDD model have similar correlation with DelPhi. For sidechain/sidechain screened Coulombic energy, there seem to be different trends between the GB models and the DDD model: while both GB models underestimate the magnitude of screened Coulombic interactions, the DDD model has inaccuracy in both directions. The large screening energy predicted by the GB methods has been attributed to the problem of micro-dielectrics in the solute interior, causing underestimation of Born radii and thus large screening energies.

The accuracy of the GB model has been shown to be highly correlated with accuracy of the Born radii.[18] There are two steps in obtaining the electrostatic energy by the GB model: calculation of Born radii for all partial atomic charges in the molecule followed by calculation of the electrostatic energy using the GB equation (Equation 3). As a test to separate inaccuracy in the GB equation from inaccuracy in Born radii calculation, sidechain desolvation and screening energies were calculated using the GB equation with Born radii calculated from DelPhi atomic solvation energies (Figure B-5A-C). The four outliers in Figure B-5B and six outliers in Figure B-5C are for cysteines, an amino acid that is generally not included in protein design calculations. Consistent with the results of Case and coworkers,[18] the agreement between the GB equation with perfect radii and energies calculated using DelPhi is much improved over the accuracy of the fully analytical Born radii calculation. There is a bias inherent in using DelPhi Born radii to reproduce DelPhi energies. However, the results in Figure B-5 do confirm that the GB equation (Equation 3), with accurate values for α, captures the physics of the Poisson equation (Equation 2).[12]

The correlation between DelPhi Born radii and analytically calculated Born radii (Figures B-5D and B-5E) is far from ideal (RMSD = 1.57 and 1.90 Å, R = 0.756 and 0.666 for GBSA and GB-PDA, respectively). In order to improve the Born radius calculation for the GBSA method, we obtained a separate parameter set for each amino acid type and the backbone by only using those particular atoms in the linear fitting. The results show an improved trend in sidechain desolvation (slope = 0.93 for aa-specific, 1.16 for general), but the scatter in the data is worse: for amino acid specific parameters, RMSD = 1.57 kcal/mol and R = 0.889 (Figure B-5F).

Another important metric in assessing the utility of an energy function is its computational efficiency. The two GB methods discussed here have comparable computational speeds. Calculation of the exact and one- and two-body energy terms for a protein with 61 amino acids (40 of which are polar) takes 22 CPU minutes on a 195 MHz SGI R10,000 processor. The same calculation using DelPhi (0.5 Å grid spacing) takes 31 CPU hours. For a typical protein design calculation there are millions of pairs. FDPB calculations for scoring rotamer pairs could potentially take 45 CPU years for 10 million pairs, while GB calculations would take 0.5 CPU years on a single processor.

Pokala and Handel[36] have reported a generic sidechain approach for implementing the GB model in protein design calculations. Their approach involves placing a sphere at a set distance from the alpha carbons for all amino acid positions other than that of the sidechain for which the Born radii are being calculated. The Born radii of the backbone atoms are calculated with the wild-type sidechains present. The GBSA method with an additional scaling factor to account for the presence of generic sidechains is used for Born radii calculations. For the sake of computational efficiency, there are no two-body

perturbations in their method, and as such, the calculation time scales linearly with protein size. The agreement with FDPB calculations for sidechain solvation energy (desolvation energy is not reported) using Pokala and Handel's generic sidechain method is better than FDPB agreement using the two-body simplified surface method for GBSA (respectively, RMSD = 0.68 and 1.14 kcal/mol, R = 0.996 and 0.995). Looking at the exact GBSA and GB-PDA sidechain solvation energy correlation with DelPhi energies (respectively, RMSD = 1.02 and 1.27 kcal/mol, R = 0.996 and 0.994), we see that the Pokala and Handel method with generic sidechains is even more accurate than the exact GB methods, as implemented here. Based on the preliminary results, it is therefore difficult to judge whether the additive generic sidechain GBSA method is a better method for decomposing solvation energy than pairwise simplified surfaces. The difference in error may be a function of the protein test set and indicate that a direct comparison between the accuracy of these two methods is only possible with the same set of molecules.

Based on our results, the GB model for polar solvation is slightly more decomposable by residue than the FDPB solver DelPhi. The GB-PDA model shows the best decomposability. However, the exact backbone and sidechain desolvation and sidechain/backbone and sidechain/sidechain screened Coulombic energies do not match DelPhi energies significantly better than currently implemented fast solvation models. In contrast to the DDD method's bias, the GB models tend to overestimate the effect of solvent screening.

References

1. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci*. **5**, 895–903.

2. Dill, K. A. (1990). Dominant forces in protein folding. *Biochem*. **29**, 7133–7155.

3. Honig, B. & Nicholls, A. (1995). Classical Electrostatics in Biology and Chemistry. *Science* **268**, 1144–1149.

4. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199–203.

5. Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA* **84**, 3086–3090.

6. Orozco, M. & Luque, F. J. (2000). Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev*. **100**, 4187–4225.

7. Sharp, K. A. & Honig, B. (1990). Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem*. **19**, 301–332.

8. Simonson, T. (2001). Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol*. **11**, 243–252.

9. Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A. & Brooks, C. L. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem*. **25**, 265–284.

10. Gilson, M. K., Sharp, K. & Honig, B. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem*. **9**, 327–335.

11. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc*. **112**, 6127–6129.

12. Bashford, D. & Case, D. A. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem*. **51**, 129–152.

13. Dominy, B. N. & Brooks, C. L. (1999). Development of a generalized born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **103**, 3765–3773.

14. Tsui, V. & Case, D. A. (2000). Theory and applications of the generalized Born solvation model in macromolecular Simulations. *Biopolymers* **56**, 275–291.

15. Tsui, V. & Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J. Am. Chem. Soc.* **122**, 2489–2498.

16. Calimet, N., Schaefer, M. & Simonson, T. (2001). Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins* **45**, 144–158.

17. Born, M. (1920). *Z. Phys.* **1**, 45–48.

18. Onufriev, A., Case, D. A. & Bashford, D. (2002). Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **23**, 1297–1304.

19. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246**, 122–129.

20. Ghosh, A., Rapp, C. S. & Friesner, R. A. (1998). Generalized born model based on a surface integral formulation. *J. Phys. Chem. B* **102**, 10983–10990.

21. Qiu, D., Shenkin, P. S., Hollinger, F. P. & Still, W. C. (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **101**, 3005–3014.

22. Lee, M. S., Feig, M., Salsbury, F. R. & Brooks, C. L. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized born calculations. *J. Comput. Chem.* **24**, 1348–1356.

23. Lee, M. S., Salsbury, F. R. & Brooks, C. L. (2002). Novel generalized Born methods. *J. Chem. Phys.* **116**, 10606–10614.

24. Schaefer, M. & Karplus, M. (1996). A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **100**, 1578–1599.

25. Schaefer, M. & Froemmel, C. (1990). A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.* **216**, 1045–1066.

26. Grycuk, T. (2003). Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation. *J. Chem. Phys.* **119**, 4817–4826.

27. Gilson, M. K. & Honig, B. (1991). The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J. Comp. Mol. Des.* **5**, 5–20.

28. Hawkins, G. D., Cramer, C. J. & Truhlar, D. G. (1996). Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges for a dielectric medium. *J. Phys. Chem.* **100**, 19824–19839.

29. Onufriev, A., Bashford, D. & Case, D. A. (2000). Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **104**, 3712–3720.

30. Sitkoff, D., Sharp, K. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978–1988.

31. Lovell, S. C., Davis, I. W., Arendall III, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Structure validation by C-alpha geometry: phi, psi, and C-beta deviation. *Proteins* **50**, 437–450.

32. Onufriev, A., Bashford, D. & Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**, 383–394.

33. Srinivasan, J., Trevathan, M. W., Beroza, P. & Case, D. A. (1999). Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **101**, 426–434.

34. Bondi, A. (1964). van der Waals Volumes and Radii. *J. Phys. Chem.* **68**, 441–451.

35. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133–152.

36. Pokala, N. & Handel, T. M. (2004). Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Sci.* **13**, 925–936.

Table B-1: **Pairwise decomposability of solvation models**

		RMSD (kcal/mol)			R		
		GBSA	GB-PDA	DelPhi	GBSA	GB-PDA	DelPhi
Backbone desolvation	*1-body*	47.2	13.3	3.03	0.998	0.999	0.998
Sidechain desolvation	*1-body*	2.41	1.38	1.91	0.868	0.840	0.746
	2-body	0.34	0.09	0.60	0.993	0.999	0.967
Sidechain/backbone screened Coulombic energy	*1-body*	1.25	0.64	1.23	0.936	0.984	0.961
	2-body	0.63	0.16	0.49	0.967	0.998	0.987
Sidechain/sidechain screened Coulombic energy	*2-body*	0.08	0.05	0.14	0.973	0.994	0.953

* All values RMSD and R values are in relation to the energy calculation with an exact dielectric boundary for the respective model.

Table B-2: **Accuracy of analytical methods compared to DelPhi***

	RMSD (kcal/mol)				R			
	GBSA	GB-PDA	LK	DDD**	GBSA	GB-PDA	LK	DDD**
Backbone desolvation	62.1	7.50	10.9	N/A	0.966	0.988	0.965	N/A
Sidechain desolvation	0.95	0.92	0.95	N/A	0.928	0.921	0.919	N/A
Sidechain/backbone screened Coulombic energy	0.88	0.98	N/A	1.18	0.951	0.964	N/A	0.903
Sidechain/sidechain screened Coulombic energy	0.10	0.11	N/A	0.13	0.977	0.961	N/A	0.930

- All energy terms were calculated using the full representation of the dielectric boundary.

** For DDD model, sidechain/backbone dielectric = 3.5r, sidechain/sidechain dielectric = 5.4r

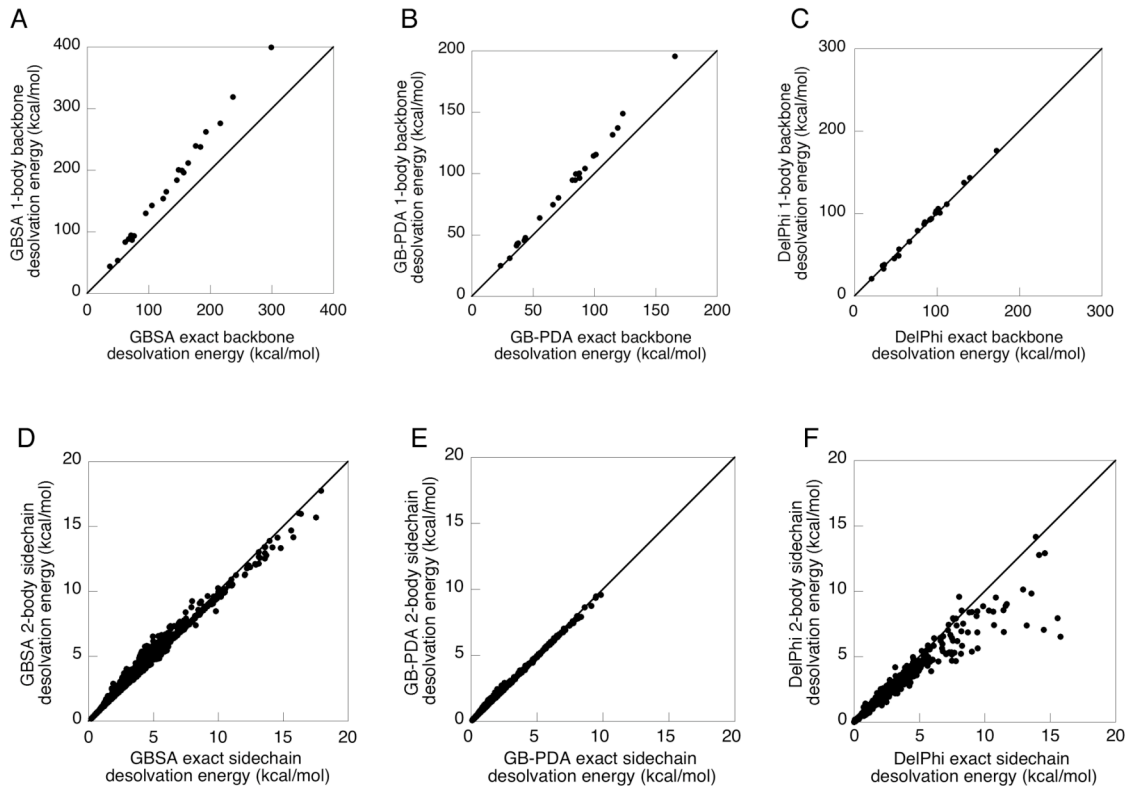


Figure B-1. One and two-body approximation for sidechain and backbone desolvation. "Exact" values refer to the energy calculated with all the wild-type amino acids used to define the dielectric boundary. Specifically for GB calculations, "exact" refers to all atoms in the protein that are included in the Born radii calculation. Backbone desolvation energy is shown in (A,B,C), and sidechain desolvation energy is shown in (D,E,F).

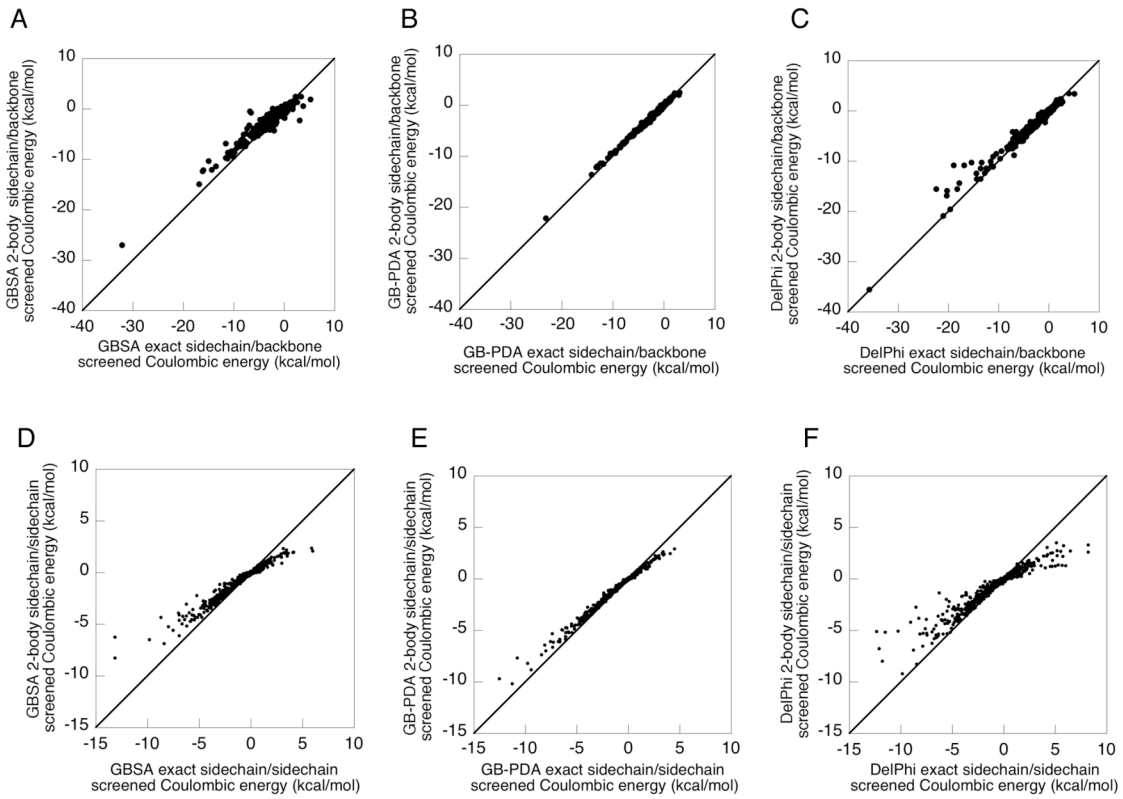


Figure B-2. Two-body decompositions for screened Coulombic energy. Sidechain/backbone screened Coulombic energy is shown in (A,B,C), and sidechain/sidechain screened Coulombic energy is shown in (D,E,F).

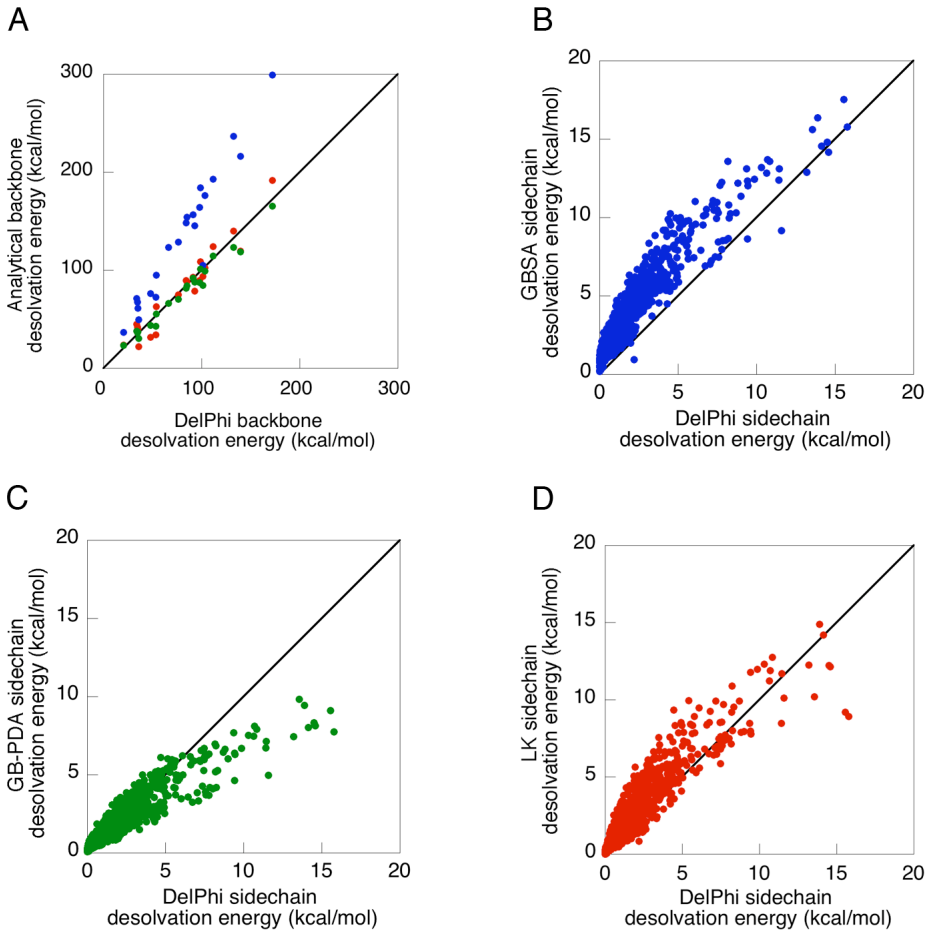


Figure B-3. Accuracy of analytical methods for calculating desolvation energy. (A) Backbone desolvation energies were calculated with DelPhi (x-axis) and analytically (y-axis). (Coloring: LK-red, GBSA-blue, GB-PDA, green). Sidechain desolvation was calculated using DelPhi (x-axis) and (B) GBSA, (C) GB-PDA, and (D) LK (y-axis).

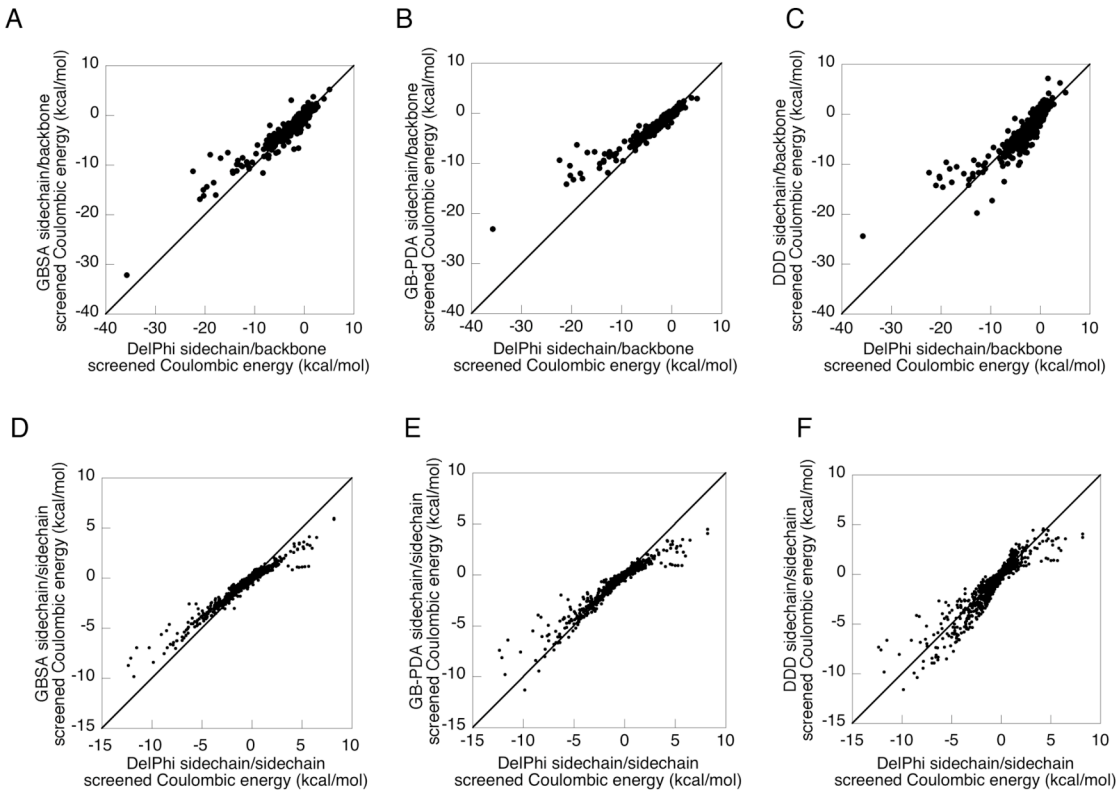


Figure B-4. Accuracy of analytical methods for calculating screened Coulombic energy. Sidechain/backbone screened Coulombic energy were calculated with DelPhi (x-axis) and with (A) GBSA, (B) GB-PDA, and (C) DDD (y-axis). Sidechain/sidechain screened Coulombic energies were calculated with DelPhi (x-axis) and with (D) GBSA, (E) GB-PDA, and (F) DDD (y-axis).

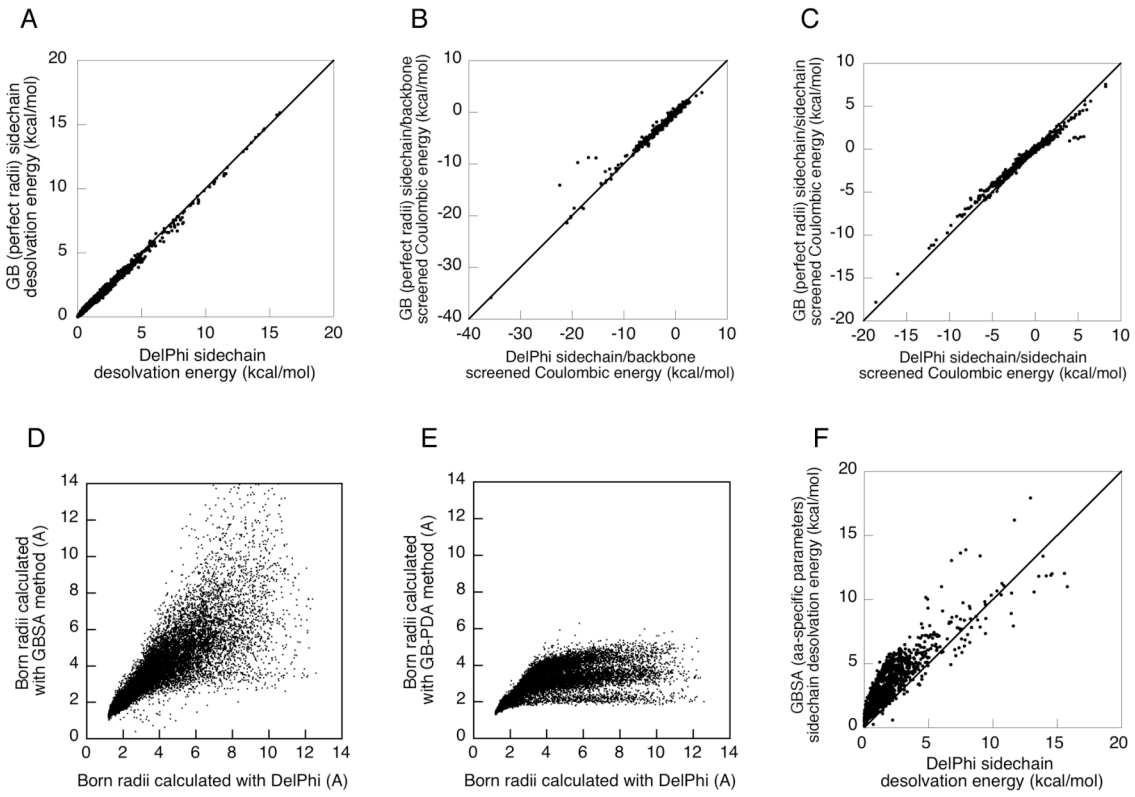


Figure B-5. The importance of Born radii. Using the GB equation (Equation 3) and Born radii from DelPhi, I calculated (A) sidechain desolvation (RMSD = 0.18, R = 0.971), (B) sidechain/backbone screened Coulombic energy (RMSD = 0.27, R = 0.964 without outliers), and (C) sidechain/sidechain screened Coulombic energy (RMSD = 0.056, R = 0.882 without outliers). Analytical Born radii were calculated using the (D) GBSA and (E) GB-PDA methods. (F) Sidechain desolvation energy was calculated using GBSA with amino acid specific parameters.

Appendix C

Designed combinatorial libraries of cytochrome p450

This project is in collaboration with Prof. Frances Arnold's group. Mike Chen collected the experimental results in the Arnold lab. Chris Snow made the ROSETTA designs.

Abstract

Cytochromes P450 represent a promising class of enzymes for engineering novel biotransformations. Our goal is to engineer the P450 BM3 from *Bacillus megaterium* to hydroxylate short-chain alkanes, specifically ethane and methane. Using two complementary strategies, we have computationally designed libraries of BM3 with mutations in the substrate-binding pocket. Experimental characterization of these libraries showed that they contain a high fraction of folded members and that at least one variant from the library with a high mutation level was active on ethane.

Introduction

Cytochrome P450 enzymes are a diverse superfamily of heme-containing monooxygenases.[1] They are crucial for a number of processes including drug metabolism and natural product synthesis. Engineered P450s have promise in the fields of drug discovery, bioremediation, and energy production.[1] With the goals of converting waste gases to transportable products and also furthering the understanding of C-H bond activation, there has been considerable progress in shifting the substrate profile of P450s toward short chain alkanes. The P450 from *Bacillus megaterium* (BM3) has been an attractive target for protein engineering due to its fused domain organization and high solubility in heterologous expression systems.[2-6] Engineering of the heme and reductase domains of BM3 has lead to variants that hydroxylate short chain alkanes with high efficiency.[4-6] To date, there are variants of both BM3 and the P450 Cam from *Pseudomonas putida* that are active on ethane.[5,7] The engineered ethane monooxygenases have volume-increasing mutations near the active site, for example, Ala to Val or Phe in the BM3 variant. Despite the dramatic shift from fatty acids or pericyclic substrates to ethane, a methane hydroxylating enzyme has not been successfully engineered.

We used computational design tools[8,9] to generate libraries of BM3 with a high level of mutation in the substrate binding pocket. Our strategy was to remove the fatty acid substrate from the crystal structure and find sequences that would fill the binding pocket while packing in energetically favorable conformations. We chose two complementary approaches: (1) the ORBIT method ensured that nearly all sequences in the library had a favorable conformation within the crystallographic backbone and (2) the CRAM method found large single mutations that could be tolerated by the structure and

built a library from these mutations with the idea that backbone relaxation would relieve steric clashes between large sidechains. Since we reasoned that a large number of mutations might be necessary to shift the native substrate profile so dramatically, we designed libraries with two possible mutations at each of ten positions, providing an average mutation level of 5 for the ORBIT library and 7.5 for the CRAM library. Data is presented here for the computational and experimental characterization of these two BM3 libraries.

Methods

The crystal structure of the BM3 heme domain with bound N-palmitoyl glycine (pdb code: 1JPZ, chain B, water removed) was subjected to 50 steps of conjugate gradient minimization using the DREIDING force field.[10] The substrate was removed after minimization. All amino acids except for Cys, Met, and Pro were allowed at nine residues: 74, 75, 78, 82, 181, 184, 188, 328, 330. Residue 87 was constrained to be either Ala or Phe in order to have half the library be in the peroxygenase family.[3] A shell of residues with sidechain atoms within 4 Å of the ten design positions was allowed to change conformation but not amino acid identity: 20, 25, 69, 71, 72, 73, 77, 81, 88, 177, 180, 185, 189, 205, 259, 260, 263, 264, 267, 268, 329, 354, 356, 436, 437, 438. A backbone independent conformer library with binning level of 1.0 was used.[11] The energy function was the same as in Treynor *et al*.[9] and Chapter 7 of this thesis. The FASTER algorithm was used to find the optimal sequence. Monte Carlo sampling with 100 temperature cycles between 150 K and 4000 K and 10^6 steps per cycle was carried out starting from the optimal sequence. The 20,000 top-scoring sequences were used to

generate a frequency table from which the amino acids for the ORBIT library were selected.[8,9] The ORBIT library composition was constrained to include the WT sequence and the most frequent amino acid contained in a degenerate codon with the WT. The "CRAM" library was designed by using the ROSETTA program to place volume-increasing amino acids at each of the nine design positions. The conformations of the single mutants were optimized using a hybrid rotamer placement and continuous minimization algorithm. A list of tolerated amino acids was generated from which the final library was chosen. Residues 74, 78, 82, 328, and 330 were allowed to mutate away from WT completely, while 75 was forced to keep the WT amino acid due to proximity with the heme moiety.

Results & Discussion

The computationally designed libraries were screened in the laboratory for their ability to bind heme and to hydroxylate a panel of substrates. Each library was constructed in two mutational backgrounds: the WT BM-3 sequence and a variant, 9-10A, that contains a number of non-active site mutations and increased activity toward propane.[4] The activity on a sampling of substrates is shown in Table C-2. The fraction folded was approximated as the fraction of library members that bound carbon monoxide and therefore heme.[12] The ORBIT libraries had a moderately higher fraction of folded variants than the CRAM library but lower activity on most substrates tested. No variant from the ORBIT library had significant ethane hydroxylating activity. Several members of the CRAM library ethane activity, the most active of which is a sevenfold mutant of

WT that supports 1800 total turnovers of ethane to ethanol: A74L/V78I/A82L/A184V/L188W/A328F/A330W.

Both libraries were screened computationally to exam the nature of their mutations. The conformation of each of 1024 sequences in each library was predicted using the ORBIT program with the same rotamer library used to design the ORBIT library. The energies of those sequences threaded onto the BM3 backbone are shown in Figure C-2A. For 25% of the CRAM library sequences, the ORBIT energy function and rotamer library were unable to find conformations with energies less than 400 ORBIT units, indicating that there are large clashes between their sidechains. The distribution of channel volume excluded by protein sidechains shows that the mutations in the CRAM library reduce the channel volume more severely than those in the ORBIT library (Figure C-2B).

We sequenced the 75 most active variants on the surrogate substrate DME.[4] The alignment of these sequences is shown in Table C-3. The majority of the highly active variants were from the CRAM libraries. Residues 74, 78, 82, 184, 188, 328, and 330 all tolerated volume-increasing amino substitutions, while residues 75, 87, and 181 were highly biased toward the WT amino acid. The consensus sequence in Table C-3 will inform further rounds of BM3 engineering. The activity data for these two libraries indicate that a more "aggressive" approach to BM3 active site mutagenesis is preferable over the more "conservative" fixed-backbone ORBIT design for generating enzymes with activity on short chain alkanes.

References

1. Urlacher, V. B. & Eiben, S. (2006). Cytochrome P450 monooxygenases: perspectives for synthetic application. *Trends Biotechnol*. **24**, 324–330.

2. Cirino, P. C. & Arnold, F. H. (2002). Protein engineering of oxygenases for biocatalysis. *Curr. Opin. Chem. Biol*. **6**, 130–135.

3. Cirino, P. C. & Arnold, F. H. (2003). A self-sufficient peroxide-driven hydroxylation biocatalyst. *Angew. Chem. Int. Ed. Engl*. **42**, 3299–3301.

4. Peters, M. W., Meinhold, P., Glieder, A. & Arnold, F. H. (2003). Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *J. Am. Chem. Soc*. **125**, 13442–13450.

5. Meinhold, P., Peters, M. W., Chen, M. M., Takahashi, K. & Arnold, F. H. (2005). Direct conversion of ethane to ethanol by engineered cytochrome P450 BM3. *Chembiochem* **6**, 1765–1768.

6. Fasan, R., Chen, M. M., Crook, N. C. & Arnold, F. H. (2007). Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting nativelike catalytic properties. *Angew. Chem. Int. Ed. Engl*. **46**, 8414–8418.

7. Xu, F., Bell, S. G., Lednik, J., Insley, A., Rao, Z. & Wong, L. L. (2005). The heme monooxygenase cytochrome P450cam can be engineered to oxidize ethane to ethanol. *Angew. Chem. Int. Ed. Engl*. **44**, 4029–4032.

8. Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A. & Dahiyat, B. I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. USA* **99**, 15926–15931.

9. Treynor, T. P., Vizcarra, C. L., Nedelcu, D. & Mayo, S. L. (2007). Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc. Natl. Acad. Sci. USA* **104**, 48–53.

10. Mayo, S. L., Olafson, B. D. & Goddard, W. A. (1990). Dreiding — a Generic Force-Field for Molecular Simulations. *J. Phys. Chem*. **94**, 8897–8909.

11. Lassila, J. K., Privett, H. K., Allen, B. D. & Mayo, S. L. (2006). Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **103**, 16710–16715.

12. Salazar, O., Cirino, P. C. & Arnold, F. H. (2003). Thermostabilization of a cytochrome p450 peroxygenase. *Chembiochem* **4**, 891–893.

Table C-1: Designed BM3 libraries

residue*	ORBIT	CRAM
A74	AV	LW
L75	LF	LF
V78	VL	FI
A82	AS	LV
F87	FA	FA
L181	LF	LW
A184	AT	AV
L188	LW	LW
A328	AF	FV
A330	AV	LW

* The WT amino acid is shown for each residue.

Table C-2: BM3 library screening*

	ORBIT		CRAM	
	WT	*9-10A*	*WT*	*9-10A*
Folded†	86	82	80	68
DME	34	54	49	53
methanol	4	5	4	6
caffeine	3	2	18	6
indole	50	67	68	79

* The number given is the percentage of library members with activity on the specified substrate. Between 700 and 900 clones were sampled for each screen.

† Folded sequences were counted as those that were able to bind carbon monoxide and therefore coordinate the heme cofactor properly.

Table C-3: Sequences of the variants with DME activity

	74	75	78	82	87	181	184	188	328	330
wt:	**A**	**L**	**V**	**A**	**F**	**L**	**A**	**L**	**A**	**A**
**						F		W	F	V
ORBIT library, 9-10A background			L	S			T	W	F	
			L	S			T		F	
			L	S		F		W	F	
			L	S				W	F	
				S			T		F	V
		F	L	S					F	V
			L	S		F			F	
				S			T	W	F	
			L	S		F			F	V
	V		L	S			T	W	F	V
	V		L	S		F	T	W	F	V
	V					F	T		F	V
CRAM library, WT background	L		I	L			V	W	F	W
	L		I	L			V	W	F	W
	L		I	L		W		W	V	W
	L		I	L			V	W	V	W
	L		F	V			V	W	F	W
	L		I	L			V	W	V	W
	L		I	L				W	V	L
	L		I	L				W	F	W
	L	F	I	V	A		V	S	L	F
	W		I	L					F	W
	W		I	L			V	W	F	W
	W		I	L			V	W	V	W
	W		I	L			V	W	V	W
	W	F	I	L			V	W	F	L
	W		F	L			V	W	F	L
	W		I	L		W		W	V	W
	W		I	L				W	F	L
	W		F	L				W	V	W
	W		F	L			V	W	F	L
	W		I	L				W	F	W
	W		I	L			V	W	F	L
	W		I	L					F	W
	W	F	I	L	A			W	F	W
	W		I	L			V	W	V	L
CRAM library, 9-10A background	L		I	L					V	W
	L		I	L			V	W	F	L
	L		I	L			V	W	F	L
	L	F	I	L			V	W	F	W
	L		I	V	A			W	F	W
	L		I	L				W	V	L
	L		I	L				W	F	W
	L		I	L			V		V	W
	L		I	L			V	W	F	L
	L		F	V				W	F	W
	L		F	L			V	W	F	W
	W		I	L					V	L
	W		F	L			V		V	L
	W		F	L			V	W	V	L
	W		I	L					V	W
	W		F	L			V		F	W
	W		F	L					F	W
	W		F	L			V		F	W
	W	F	I	L					V	L
	W		I	V			V		F	W
	W		I	V				W	F	W
	W	F	I	L	A			W	V	W
	W	F	F	V	A				V	L
	W		F	L			V		F	W
	W		I	V	V				F	W
	W		I	L			V	W	F	W
	W		I	L			V		V	W
	W		I	L				W	F	W
	W		I	L				W	V	W
	W		F	L			V		V	W
	W		F	L			V		F	W
	W		F	L					F	W
	W	F	I	L			V	W	F	W
	W		F	L			V		F	W
consensus:	LW	-	ILF	LS	-	-	V	LW	FV	W

* Blank lines correspond to WT amino acid identity.

** ORBIT library with WT background (one sequence)

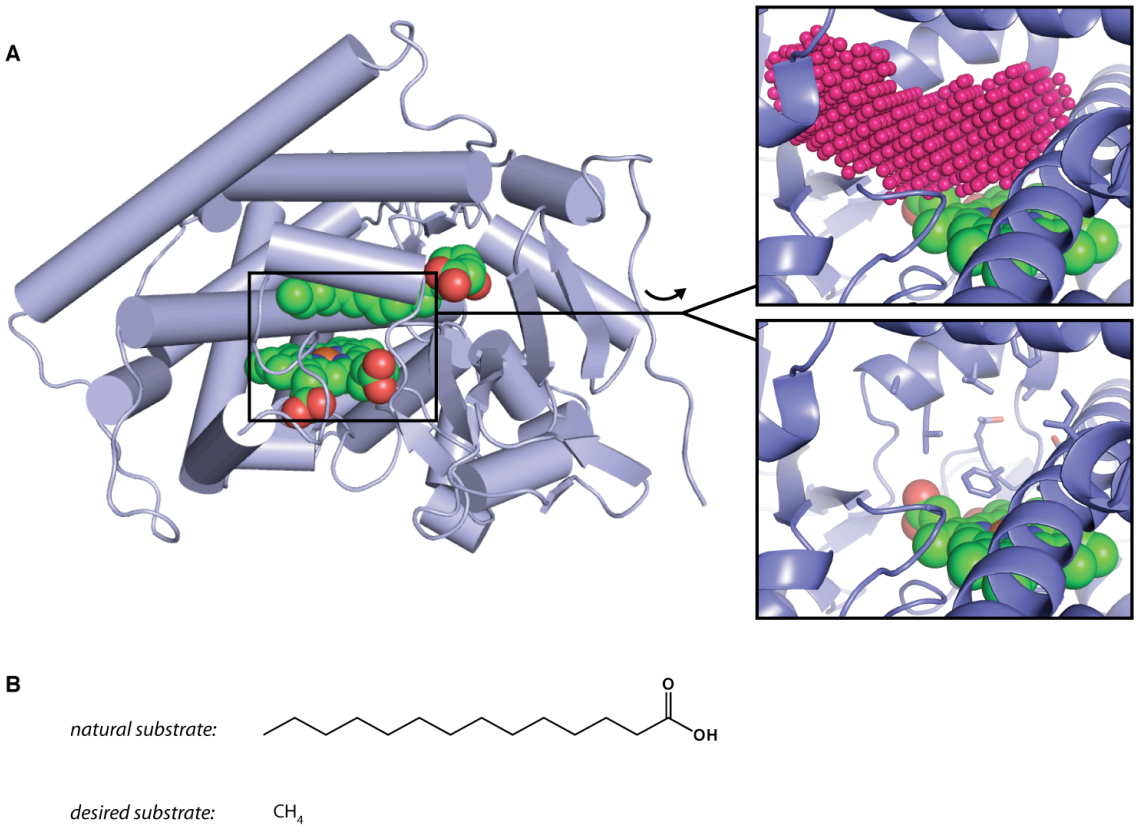


Figure C-1. The BM3 structure. (A) The heme domain of BM3 is shown in cartoon representation (pdb code: 1JPZ). The heme and N-palmitoyl glycine are shown as spheres. The magnified images show the rotated substrate-binding pocket with the substrate removed. The top panel shows the cavity left by removing the substrate molecule as a grid of pink spheres. In the bottom panel, the residues targeted for mutagenesis are shown as sticks. (B) The WT BM3 protein hydroxylates C12–C18 fatty acids at sub-terminal positions. Myristic acid is shown here. The desired substrate for this engineering project is methane.

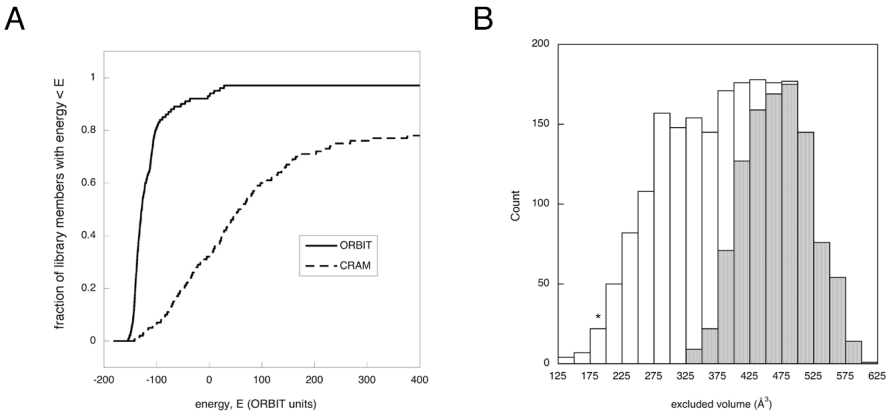


Figure C-2. Properties of the designed libraries. (A) Cumulative histogram of library energies. The ORBIT energies were calculated for the predicted conformations of each sequence in the ORBIT and CRAM libraries. (B) Volume histograms for the ORBIT (white) and CRAM (gray) libraries. In the minimized crystal structure (with the substrate removed), the ten design positions were replaced with Ala. A grid of points was defined at locations that were not occupied by any protein atoms in the Ala-substituted structure. The predicted conformations of the designed proteins were scored according to how many grid points overlapped with their atomic radii (excluded volume). The excluded volume of the WT crystallographic rotamers is marked with a star on the plot. For both plots, the structures of the variants were modeled in the background of the WT sequence.